

Copyright  
by  
Jianchao Yao  
2009

**The Dissertation Committee for Jianchao Yao Certifies that this is the approved  
version of the following dissertation:**

**Integrative Analysis of High-Throughput Biological Data: Shrinkage  
Correlation Coefficient and Comparative Expression Analysis**

**Committee:**

---

Stanley J. Roux, Supervisor

---

Zengjian J. Chen, Co-Supervisor

---

Mia K. Markey

---

Daniel P. Miranker

---

Bo Fu

**Integrative Analysis of High-Throughput Biological Data: Shrinkage  
Correlation Coefficient and Comparative Expression Analysis**

**by**

**Jianchao Yao, B.S.; M.S.**

**Dissertation**

Presented to the Faculty of the Graduate School of  
The University of Texas at Austin  
in Partial Fulfillment  
of the Requirements  
for the Degree of

**Doctor of Philosophy**

**The University of Texas at Austin  
December, 2009**

## **Dedication**

To my dear parents and sisters

## **Acknowledgements**

I would like to express my deep and sincere gratitude to my supervisor, Dr. Stan Roux, for his continuous encouragement, optimism, support and guidance. I want to thank him for providing me several opportunities in the years I have known him. I have learned from him the value of careful thought, precise word selection, and ambassadorial treatment of colleagues.

I also want to thank Dr. Mari Salmi, Dr. Jian Wu, Dr. Chunqi Chang, Dr. Ann Loraine, and Dr. Y.S. Hung, for their support and collaboration over the years. I would like to thank Dr. Vishy Iyer for introducing me to genomics; Dr. Enamul Huq, Dr. Mona Mehdy, and all of our lab members for insightful discussions; and Dr. Greg Clark for constant encouragement. I thank all my friends who have stuck by me over the years and encouraged me. I am grateful to my committee members for their time, efforts, and concern.

Lastly, I would like to thank my family for their love, understanding, patience, and support. They encouraged me at all times, especially when I needed it most.

# **Integrative Analysis of High-Throughput Biological Data: Shrinkage Correlation Coefficient and Comparative Expression Analysis**

Publication No. \_\_\_\_\_

Jianchao Yao, Ph.D.

The University of Texas at Austin, 2009

Supervisor: Stanley J. Roux

Co-Supervisor: Zengjian J. Chen

The focus for this research is to develop and apply statistical methods to analyze and interpret high-throughput biological data. We developed a novel correlation coefficient, shrinkage correlation coefficient (SCC), that fully exploits the similarity between the replicated microarray experimental samples. The methodology considers both the number of replicates and the variance within each experimental group in clustering expression data, and provides a robust statistical estimation of the error of replicated microarray data. Applying SCC-based hierarchical clustering to the replicated microarray data obtained from germinating spores of the fern *Ceratopteris richardii*, we discovered two clusters of genes with shared expression patterns during spore germination. This computational approach is not only applicable to DNA microarray analysis but is also applicable to proteomics data or any other high-throughput analysis methodology.

The suppression of *APY1* and *APY2* in mutants expressing an inducible RNAi system resulted in plants with a dwarf phenotype and disrupted auxin distribution, and we used these mutants to discover what genes changed expression during growth suppression. We evaluated the gene expression changes of apyrase-suppressed RNAi mutants that had been grown in the light and in the darkness, using the NimbleGen *Arabidopsis thaliana* 4-Plex microarray, respectively. We compared the two sets of large-scale expression data and identified genes whose expression significantly changed after apyrase suppression in light and darkness, respectively. Our results allowed us to highlight some of the genes likely to play major roles in mediating the growth changes that happen when plants drastically reduce their production of APY1 and APY2, some more associated with growth promotion and others, such as stress-induced genes, more associated with growth inhibition. There is a strong rationale for ranking all these genes as prime candidates for mediating the inhibitory growth effects of suppressing apyrase expression, thus the NimbleGen data will serve as a catalyst and valuable guide to the subsequent physiological and molecular experiments that will be needed to clarify the network of gene expression changes that accompany growth inhibition.

## Table of Contents

List of Tables .....	x
List of Figures .....	xi
Chapter 1 Introduction .....	1
Microarray Technologies .....	1
Two-Channel Microarrays .....	2
One-Channel Microarrays .....	2
Microarray Data Analysis .....	3
Preprocessing The Data .....	3
Measuring Similarity of Expression Patterns .....	5
Visualizing Microarray Data .....	6
Cluster Analysis of Microarray Data .....	7
Identifying Differentially Expressed Genes .....	8
Biological Interpretation of Microarray Data .....	10
RNA-Sequencing .....	11
RNA Library Construction .....	12
Next-Generation Sequencing .....	12
Data Analysis .....	13
Research Outline .....	16
Chapter 2 Genome-Scale Analysis of <i>Ceratopteris richardii</i> cDNA Microarray Data Using Shrinkage Correlation Coefficient .....	18
Background .....	18
Methods .....	23
Data Source, Retrieval, and Missing Value Imputation .....	23
Singular Value Decomposition .....	24
Shrinkage Correlation Coefficient .....	25
Gene Ontology/Functional Enrichment Analysis .....	37
Results .....	38
<i>C. richardii</i> Microarray Data Quality Control .....	38



Cluster Analysis of <i>C. richardii</i> Gene Expression.....	44
Discussion .....	55
Shrinkage Correlation Coefficient is a Robust Correlation .....	55
Results of Functional Analysis .....	57
Chapter 3 Comparative Analysis of <i>Arabidopsis thaliana</i> Microarray Data .....	62
Background .....	62
Methods.....	63
Plant Materials and Growth Conditions.....	63
Total RNA Isolation and NimbleGen Microarray Experiments .....	63
False Discovery Rate .....	63
Significance Analysis of Microarrays.....	65
Results.....	66
Discussion .....	80
Chapter 4 Conclusions and Future Direction.....	83
Future Directions .....	84
Bibliography .....	85
Vita .....	93

## List of Tables

Table 2.1:	GO categories significantly ( $FDR < 0.10$ ) enriched among genes belonging to SCC Clusters A and B .....	53
Table 3.1:	33 genes upregulated both in light- and dark-grown DKO mutants .....	73
Table 3.2:	45 genes down-regulated both in light- and dark-grown DKO mutants .....	74
Table 3.3:	Five up-regulated genes are stress related and three down-regulated genes are growth-promoting genes .....	77
Table 3.4:	Real-Time RT-PCR verification of microarray expression .....	79

## List of Figures

Figure 2.1: The performance of the three models indicated by the adjusted Rand index obtained from the synthetic data sets using hierarchical clustering and k-means clustering.....	33
Figure 2.2: The performance of the three correlations indicated by the adjusted Rand index obtained from the real yeast expression data using hierarchical clustering and k-means clustering.....	36
Figure 2.3: Identification of low-quality arrays .....	39
Figure 2.4: Gene-wise bias (Eigengene 5) associated with the two prints of arrays .....	47
Figure 2.5: Histogram of optimal shrinkage factor $\lambda_i^*$ .....	49
Figure 2.6: TUG expression profile in the early stages of gametophyte development of <i>C. richardii</i> by SCC Identification of low-quality arrays .....	50
Figure 3.1: Number of genes altered by DKO treatment in light-grown and dark-grown seedlings .....	68
Figure 3.2: The number of up- or down-regulated genes altered by the DKO treatment in light- and dark-grown seedlings .....	70
Figure 3.3: Functional categorization of the up-regulated 33 genes expressed both in light- and dark-grown seedlings by the DKO treatment.....	71
Figure 3.4: Functional categorization of the down-regulated 45 genes expressed both in light- and dark-grown seedlings by the DKO treatment.....	72

## **Chapter 1 Introduction**

### **BIOINFORMATICS**

Bioinformatics is a fusion of biology, statistics, and engineering. It involves using statistical and computational methods to address research questions in biology. Since the late 1990s, the discipline of bioinformatics has experienced tremendous growth, thanks mainly to the success of large-scale projects such as the Human Genome Project and related efforts which generate vast amounts of data. Many research efforts in this field include genomics (e.g., gene expression analysis, network analysis), proteomics (e.g., protein-protein interaction, prediction of protein structure), and metabolomics (e.g., nuclear magnetic resonance spectroscopy). In this dissertation, we will focus on the use of microarray for gene expression analysis.

### **MICROARRAY TECHNOLOGIES**

Since the mid-1990s, the amount of data produced by molecular biologists has been growing at an exponential rate. Some of the fastest growing sets of data are measurements of gene expression (i.e., the transcription of the genetic information contained within the DNA into messenger RNA molecules that are then translated into the proteins that perform most of the critical functions of cells), such as large-scale data sets generated by microarray technology. As the next revolution in molecular biology, microarray technology enables scientists to examine the expression levels of thousands of genes at the same time. These gene expression data offer hints as to the functions of thousands of newly discovered genes.

## **Two-Channel Microarrays**

For two-channel microarray experiments, cDNA preparations are generated from two samples (e.g. treatment and control) and then labeled with two different fluorescent dyes (e.g. Cy5-red and Cy3-green). The two Cy-labeled cDNA samples are then mixed together and hybridized to a single microarray. After hybridization, the microarray is scanned in a scanner to visualize fluorescence of the two dyes, and the relative intensities of each dye are used in ratio-based analysis to identify up-regulated and down-regulated genes. An example of the two-channel system is the cDNA microarrays discussed in Chapter 2.

## **One-Channel Microarrays**

In one-channel microarrays, the absolute values of gene expression are estimated. A single dye is used for all samples. Therefore, the comparison of two conditions requires two separate single-dye hybridizations. One advantage of the one-channel microarrays is that a low-quality sample cannot affect the raw data generated from other samples, because each sample is hybridized to a single microarray. Another benefit is that data can be easily compared between different experiments because each gene expression level is an absolute value. One drawback of the one-channel microarrays is that twice as many microarrays are needed in one set of experiments compared to the two-channel microarrays. Two popular one-channel systems are the NimbleGen microarrays and the Affymetrix “Gene Chip”. NimbleGen microarrays are discussed in Chapter 3.

## MICROARRAY DATA ANALYSIS

### Preprocessing The Data

Before microarray data are stored in a database or analyzed in various ways, a number of transformations may have been done to it which include calculating ratios of the raw data, log-transforming these ratios, zero-centering a gene or a sample expression pattern, and imputing missing data. Furthermore, all the transformations described below can be applied to data from any microarray platform.

*Deriving ratios from the raw data.* Most microarray experiments look for genes that are differentially expressed. Assuming that we have a total number of  $N$  genes in an array, then the ratio of the  $i$ th gene (i.e., expression change) can be written as

$$T_i = \frac{R_i}{G_i}$$

where  $R$  and  $G$  represent a treatment and a control sample respectively, and  $i$  is an index running over all the genes from 1 to  $N$ .

Although ratios provide an intuitive measure of expression changes, one disadvantage with ratios is that they treat up- and down-regulated genes differently. Genes upregulated by 2-fold have an expression ratio of 2, whereas those downregulated by the same fold have an expression ratio of 0.5.

*Log-transforming the ratio data.* The logarithm base 2 is the most widely used transformation, which produces ratios that are often easy to interpret by biologists. Logarithms treat numbers and their reciprocals symmetrically, for example,  $\log_2(2) = 1$ ,  $\log_2(1/2) = -1$ . The logarithms of the expression ratios are also treated symmetrically. Genes upregulated by 2-fold have an  $\log_2(\text{expression ratio})$  of 1, whereas those downregulated by the same fold have an  $\log_2(\text{expression ratio})$  of -1.

*Centering data.* The following transformation is usually used in the context of the log transformed data.

$$X' = X - \bar{X}$$

where  $X$  is the  $\log_2$  – transformed expression pattern of a certain gene (i.e., a row vector across all the arrays) and  $\bar{X}$  is the arithmetic mean of all the components in this row vector. Therefore, centering sets the average value of the expression pattern of a certain gene to zero, i.e.,  $\bar{X}' = 0$ . Centering data can be used to compare expression patterns of different genes. It is useful when the actual value of the expression ratio is not important or is not meaningful (e.g., common reference). However, it is generally not appropriate when using a biologically meaningful control sample, such as a matched, untreated sample, or a zero time point.

*Missing value imputation.* Microarray experiments often generate data sets with a number of missing values which are caused by various reasons, including insufficient dye on certain spots, image corruption, or simply dust on the slide. Missing data can be imputed by some simple approaches, for example, flagging their positions manually and excluding them from subsequent analysis, or replacing missing log-transformed values by zero or by an average expression over the rows (samples or experiments). Such approaches are not optimal because they do not consider the correlation structure of the data. A more sophisticated method which takes advantage of the correlation structure of the data is the K-nearest neighbors algorithm (KNN) (Troyanskaya, Cantor et al. 2001). To impute missing values, the KNN method selects genes with expression patterns similar to the gene of interest. If there is a missing value in experiment  $j$  for gene  $i$ , the KNN method will select  $K$  other genes, which have expression values in experiment  $j$ , with expression patterns most similar to gene  $i$  in the remaining experiments. A weighted

average is calculated from the  $K$  closest genes and used as an estimate for the missing value in experiment  $j$  for gene  $i$ .

## Measuring Similarity of Expression Patterns

One of the fundamental processes in microarray analysis is to measure the similarity of two expression patterns. There are many different ways in which a measure of similarity can be calculated. In the next paragraphs two similarity metrics will be discussed.

*Pearson Correlation Coefficient.* The Pearson correlation of the two vectors  $x = (x_1, x_2, \dots, x_n)$  (e.g., expression pattern of gene  $x$  over  $n$  time points) and  $y = (y_1, y_2, \dots, y_n)$  (e.g., expression pattern of gene  $y$  over  $n$  time points) is:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

A Pearson correlation coefficient indicates the relationship between two ordered sets of numbers (e.g., gene or array expression patterns) and the strength of this relationship. For example, if two genes increase or decrease proportionally over time, their correlation will be 1.0. If two genes have no relationship to each other, their correlation will be 0. If two genes express in the opposite direction over time, their correlation will be -1.0. A correlation coefficient is always between -1 and 1. It is particularly useful when we look for genes with the exact same pattern even at different levels of variation.

*Euclidean distance.* When we look for genes with a profile in the same level of variation, we can use the Euclidean distance. The Euclidean distance between two  $n$ -dimensional vectors  $x = (x_1, x_2, \dots, x_n)$  (e.g., expression pattern of gene  $x$  over  $n$  time points) and  $y = (y_1, y_2, \dots, y_n)$  (e.g., expression pattern of gene  $y$  over  $n$  time points) is:



$$d_E(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

If two genes have identical profiles, the Euclidean distance between them will be zero. The larger the Euclidean distance, the less similarity is between two genes.

## Visualizing Microarray Data

Microarray data are stored in rows (genes) and columns (experiments). Visualization of microarray data is extremely important for biological knowledge discovery. There are a number of methods to visualize microarray data, ranging from simple to sophisticated.

*Scatter plot.* The scatter plot is a 2D or 3D plot in which a vector is displayed as a point having the coordinates equal to the components of the vector. It is probably the simplest tool that can be use in microarray visualization. This technique can visualize interactions between two variables. For example, in a scatter plot, each axis can correspond to a gene and each expression level corresponding to an individual experiment is represented as a point. This plot can suggest certain relationship between the expression levels of two genes.

*Time series.* A time series is a plot in which the expression values of genes are plotted against the time points when the values were measured. The horizontal axis is the time and the vertical axis represents the expression values. In the time series plot, each wave corresponds to a gene across experiments on the horizontal axis.

*Heat map.* A heat map is a graphical representation of data where the values from a data matrix are represented by colors. Biological heat maps, as presented in (Eisen, Spellman et al. 1998) are usually used to represent the expression levels of genes across

different experiments. Larger expression values are represented by small dark squares and smaller values by lighter squares.

*Dendrogram.* This is one of the most effective and powerful tools to illustrate the hierarchical view of relationships between genes or arrays. In this tree-like plot, each step of hierarchical clustering is represented as a fusion of two branches of the tree into a single one. Each branch represents a single gene or array expression.

*Principal component analysis (PCA).* In microarray expression analysis, each gene and each experiment may represent one dimension. For example, a set of 10 experiments involving 20,000 genes may be conceptualized as 20,000 data points (genes) in a 10-dimensional space, or 10 data points (experiments) in a 20,000-dimensional space. This large number of dimensions causes the complexity of data analysis. A natural approach to solve this problem is to try to reduce the number of dimensions by eliminating the less important dimensions. PCA is such an approach that keeps the dimensions that accounts for a large variance in the original data and filters out the dimensions that accounts for a small portion of the total variance of the original data. PCA (Anderson 2003) is also known as singular value decomposition (Golub and Van Loan 1996) in linear algebra. The details will be discussed in Chapter 2.

### **Cluster Analysis of Microarray Data**

The aim of cluster analysis or clustering is to partition a set of observations into different groups (clusters) so that observations in the same cluster are similar in some sense. The similarity can be measured by distance metrics, for example, Pearson correlation coefficient. Cluster analysis is the first most widely used method to analyze microarray data. It can separate genes into different clusters, and in the same cluster, genes are grouped together because they are involved in the same biological processes or

share similar molecular functions. Here, we describe two classical clustering algorithms and both of them are supported in most commercial microarray data analysis software.

*Hierarchical clustering.* The aim of hierarchical clustering is to build the hierarchy of clusters. There are two types of hierarchical clustering: agglomerative and divisive. Agglomerative hierarchical clustering is a bottom-up clustering method in which each single observation (gene or sample) starts in its own cluster. Then, the closet pair of clusters are agglomerated by satisfying some similarity criteria in each successive iteration until all the observations are in one cluster. The similarity between sets of observations can be calculated by the foregoing similarity metrics. Divisive hierarchical clustering is a top-down approach where all observations start in one cluster. Then, each cluster is subdivided into smaller pieces in each successive iteration. Divisive hierarchical clustering is not generally available, and rarely has been applied.

*K-means clustering.* K-means clustering is one of the simplest partition clustering method. The aim of k-means clustering is to classify observations into k number of clusters in which each observation is assigned to the cluster with the nearest mean. The classification is done by minimizing the sum of squares of distances between data and the centroid of the corresponding cluster.

## **Identifying Differentially Expressed Genes**

In many cases, the purpose of microarray experiment is to compare the expression levels of genes in two different specimens, for example, healthy vs. disease or treated vs. control samples. In all such comparative studies, a very important problem is to identify genes differentially expressed in the two samples compared. In the following, we will discuss several methods which can be used to distinguish genes truly differentially expressed from those that are simply affected by microarray experimental noise.

*Fold change.* The simplest and most intuitive method to find differentially expressed genes is to consider their fold change between treatment and control. Typically an arbitrary threshold such as 2 is chosen and any change with expression value larger than +2 or smaller than -2 will be considered as significant. Although the fold-change method is simple and intuitive, it has important disadvantages. One of the major disadvantages is that the threshold is chosen arbitrarily and may not be appropriate. For example, if many genes under study have dramatic expression changes, the method will select many genes and have a low specificity. In addition, the fold-change method is not a statistical test, and there is no associated level of confidence in the designation of a gene as being differentially expressed or not.

*Hypothesis testing, corrections for multiple comparisons.* Another method to select differentially expressed genes is to use the classical hypothesis testing approach (e.g., t-test) in conjunction with some correction for multiple comparisons. Let us consider an experiment in which gene expression levels are compared between tumor and healthy tissue. Assuming that we have five tumor samples, five healthy tissue samples, and 20 independent genes, we can perform a t-test gene-by-gene for means involving two conditions (i.e., tumor vs. healthy tissue). The null hypothesis will be that there is no gene expression difference between tumor and healthy tissue, and genes with p-values lower than the significance level (e.g., 5%) will be called significant or differentially expressed. However, the so-called significant genes may be there due to random factors such as noise. The genes that are called differentially expressed when in fact they are not will be false positives. The mistake made to report a true non-significant gene as a significant one (i.e., reject a true null hypothesis) is called a Type I error. When we perform a test for a single gene, the probability of a Type I error is controlled by the significance level at the gene level (single comparison). However, when we perform many statistical tests

simultaneously for genes in a high density array (multiple comparison), the significance level at individual gene level does not control the overall significance level at the experiment level (i.e., overall probability of making a Type I error or family-wise error rate) anymore. In another way, while a given significance level may be appropriate for each individual comparison, it is not for the set of all comparisons. To ensure the overall significance level, we need to lower the significance level for individual genes to account for the number of comparisons being performed (i.e., the number of genes). Bonferroni (Bonferroni 1935) and Sidak (Sidak 1967) corrections are two simple multiple-comparison correction methods. However, both of them are not suitable for gene expression analysis because for large number of genes, the required significance at the gene level becomes very small. In other words, the Bonferroni and Sidak corrections are very conservative methods in the sense that if a gene is significant after a Bonferroni or Sidak correction, then the gene is truly differentially expressed. However, if a gene is not significant after correction, it may still be truly differentially expressed. In the context of microarray data, false discovery rate (FDR) (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001) and significance analysis of microarray (SAM) (Tusher, Tibshirani et al. 2001) are suitable methods for multiple comparison corrections. These two methods will be discussed in Chapter 3.

### **Biological Interpretation of Microarray Data**

After differentially expressed genes are selected, the challenge for researchers is to interpret the microarray results and identify the underlying biological functions or mechanisms. An ontology developed by the Gene Ontology (GO) Consortium (Ashburner, Ball et al. 2000) can meet this challenge. GO describes attributes of genes, for example, biological process, molecular function and cellular component. In addition,

it can explore functional annotations of genomes of different organisms automatically. The GO database can be browsed by the AmiGO browser in tree or directed acyclic graph view. Users can search the database using a GO term or a gene product. The results of the search performed with a gene product will contain the gene and all annotation information associated to this gene or gene product.

## **RNA-SEQUENCING**

The transcriptome is the complete set of mRNA molecules (i.e., transcripts) expressed in one cell or a population of cells. Understanding the transcriptome can help researchers determine when and where a gene is turned on or off in various types of cells or tissues. In addition, the study of transcriptomics, also referred to as expression profiling, can help researchers determine the amount of gene activity (i.e., expression level) in a certain cell or tissue type.

Various high-throughput technologies have been developed to study the transcriptome, including DNA microarrays and next-generation sequencing (also called massively parallel sequencing) technologies. Apart from the foregoing advantages, DNA micorarrays have several limitations, which include: low resolution of the output; a limited dynamic range of changes that can be observed; high background noise due to hybridization. Compared to DNA microarrays, the currently emerging next-generation sequencing technologies have higher resolution of the output, and a dynamic range that is orders of magnitude greater than microarrays (Wilhelm and Landry 2009).

RNA-Sequencing (RNA-Seq) (Mortazavi, Williams et al. 2008; Nagalakshmi, Wang et al. 2008) is a recently developed method that uses next-generation sequencing technologies to profile transcriptome. It has already been applied to human and mouse

cells, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Arabidopsis thaliana* (Cloonan, Forrest et al. 2008; Lister, O'Malley et al. 2008; Marioni, Mason et al. 2008; Morin, Bainbridge et al. 2008; Mortazavi, Williams et al. 2008; Nagalakshmi, Wang et al. 2008; Wilhelm, Marguerat et al. 2008). We will discuss how RNA-Seq works in the following paragraphs.

### **RNA Library Construction**

Prior to high-throughput sequencing, we need to generate a cDNA library in which adaptors are attached to one or both ends of cDNA fragments. To create this library, we first use PolyA enrichment steps to selectively remove ribosomal RNA or selectively enrich for mRNA. Following enrichment, the PolyA<sup>+</sup> enriched mRNA will be primed for the reverse transcription reaction using either random primers or oligo dT primers. Once the cDNA is synthesized it can be further fragmented to reach the desired fragment length. Each sequence is typically 30-400 bp, depending on the following sequencing technology used.

### **Next-Generation Sequencing**

Three next-generation sequencing technologies can be used for RNA-Seq: the Applied Biosystems SOLiD<sup>TM</sup> System (Cloonan, Forrest et al. 2008) ([http://marketing.appliedbiosystems.com/images/Product/Solid\\_Knowledge/flash/102207/solid.html](http://marketing.appliedbiosystems.com/images/Product/Solid_Knowledge/flash/102207/solid.html)), the Illumina/Solexa Genome Analyzer (Bentley 2006) (<http://www.illumina.com/pages.ilmn?ID=203>), and the Roche/454 FLX (Margulies, Egholm et al. 2005) (<http://www.454.com/enabling-technology/the-system.asp>). Another two platforms have recently been developed: the Helicos Heliscope<sup>TM</sup>

([www.helicosbio.com](http://www.helicosbio.com)) and Pacific Biosciences SMRT ([www.pacificbiosciences.com](http://www.pacificbiosciences.com)).

They have not yet been used for RNA-Seq, but are also appropriate.

The foregoing sequencing technologies generate millions of short reads (i.e., DNA sequences) from a cDNA library. Then, these reads are laid out on a single chip and sequenced in parallel. The procedures used to lay out those reads on a single chip are different for various sequencing technologies: SOLiD and 454 first attach DNA to coated beads, whereas Solexa and Helicos attach DNA directly to the chip. Although these technologies differ in chip generation, they share the same sequencing mechanism: after a single cDNA fragment is isolated and attached to a solid matrix, this single molecule will be amplified either by bridge PCR (Solexa) or emulsion PCR (SOLiD/454). Then, the cDNA fragments are sequenced in parallel, either by the measurement of the incorporation of short fluorescent linkers (SOLiD) or fluorescent nucleotides (Solexa), or by the release of pyrophosphate from incorporation of normal nucleotides (454).

Other differences between these platforms are the run time required to generate data and the resulting read-length. The run times vary from 8 h to 10 days, depending on the platform and read type (single end or paired ends). Compared to the other major two platforms, the 454 platform generates smallest amount of reads (~400,000 reads) but has the longest read-length (200-300 bp). SOLiD and Solexa generate tens of millions of reads with read-length 35 to 36 bp.

## **Data Analysis**

*Data filtering.* Once generated, sequence reads will need to be mapped to the reference genome, or used for *de novo* sequence assembly to reconstruct the original sequence structures. However, as part of the results from any current sequencing run, certain percent of the short reads are low-quality sequences that cannot be mapped. They



may contain reads that span exon-exon junctions or that contain PolyA tails. As discussed above, the PolyA enrichment steps are used to selectively enrich mRNA so that splicing has taken place and the resulting short reads cannot come from intronic sequences. Therefore, when the short reads are aligned to the reference genome, only reads from within exonic regions can be mapped whereas those from exon-exon junctions cannot. Exon-exon junctions can be identified by the presence of the GT-AG dinucleotides that flank splice sites, and PolyA tails can be identified by the presence of multiple As or Ts at the end of some reads. Filtering these low-quality reads will accelerate subsequent downstream analysis.

*Read mapping.* Once the low-quality reads have been filtered out, the next challenge is to map the high-quality short reads to the reference genome. There are several programs available for mapping reads to the reference genome, including SHRiMP (Rumble, Lacroute et al. 2009) (<http://compbio.cs.toronto.edu/shrimp/>), Bowtie (Langmead, Trapnell et al. 2009) (<http://bowtie-bio.sourceforge.net/index.shtml>), Mapreads (<http://solidsoftwaretools.com/gf/project/mapreads/>), ZOOM (Lin, Zhang et al. 2008) (<http://www.bioinformaticssolutions.com/products/zoom/index.php>), SOAP (Li, Li et al. 2008) (<http://soap.genomics.org.cn/#pub2>), RMAP (Smith, Xuan et al. 2008) (<http://rulai.cshl.edu/rmap/>), MAQ (Li, Ruan et al. 2008) (<http://maq.sourceforge.net/>), ELAND (part of the Illumina suite).

*Expression measurement.* In order to derive gene expression levels, a method must be used to convert reads into a quantitative value for each gene. For RNA fragmentation followed by cDNA synthesis, we can sum the number of reads that fall into the exons of a gene, and then normalize the total number by the length of exons. And, for the 3'-biased method, we can divide the sum of reads within an arbitrary portion of the 3' end of the ORF by the same arbitrary length (Nagalakshmi, Wang et al. 2008).

Once expression levels are measured, one can visualize the data along with genome annotation information for specific regions and this illustration can facilitate the identification of novel genes.

## RESEARCH OUTLINE

The focus for this research is to develop and apply statistical methods to analyze and interpret high-throughput biological data. In Chapter 2, we describe a novel correlation coefficient, shrinkage correlation coefficient (SCC), that fully exploits the similarity between the replicated microarray experimental samples. The methodology considers both the number of replicates and the variance within each experimental group in clustering expression data, and provides a robust statistical estimation of the error of replicated microarray data. Applying SCC-based hierarchical clustering to the replicated microarray data obtained from germinating spores of the fern *Ceratopteris richardii*, we discovered two clusters of genes with shared expression patterns during spore germination. Functional analysis suggested that some of the genetic mechanisms that control germination in such diverse plant lineages as mosses and angiosperms are also conserved among ferns.

In Chapter 3, we describe and analyze microarray data that reveal the gene expression changes that accompany growth inhibition when growth-regulating enzymes called apyrases (NTPDases) are suppressed. Apyrases are enzymes that remove the terminal phosphate from nucleoside triphosphates (e.g., ATP) and nucleoside diphosphates. A recent study in our lab revealed that two apyrase genes, *AtAPY1* and *AtAPY2*, play important roles in the control of plant growth in *Arabidopsis* (Wu, Steinebrunner et al. 2007). Specifically, the suppression of these apyrases in an inducible RNAi system resulted in plants with a dwarf phenotype and disrupted auxin distribution. To better understand the implications of these findings, an analysis of the underlying gene expression changes that accompany *APY* gene suppression was carried out. We used an inducible RNAi construct to suppress *APY1* in plants homozygous for the *apy2* knockout mutation. Growth inhibition of the mutant seedlings becomes evident after 3 d

of growth in the presence of the estradiol inducer. We compared gene expression differences between uninduced plants and plants grown continuously in the inducer for 3.5 d (dark grown) or 6 d (light-grown) using the NimbleGen *Arabidopsis thaliana* 4-Plex microarray. We compared the two sets of large-scale expression data and identified genes whose expression significantly changed after ectoapyrase suppression in light- and dark-grown plants, respectively. Major changes in numerous transcription factors and in hormone-regulated genes were observed, and four of them were independently verified by qRT-PCR. Data analysis has provided a better understanding of the molecular bases underlying the relationship of ectoapyrase expression to growth. We describe the foregoing comparative expression analysis in Chapter 3. Conclusions as well as future directions are summarized in Chapter 4.

## **Chapter 2 Genome-Scale Analysis of *Ceratopteris richardii* cDNA Microarray Data Using Shrinkage Correlation Coefficient**

### **BACKGROUND**

Advances in high-throughput technologies, such as DNA microarrays and genome sequencing, have enabled the large-scale exploration of the genome in a way that is systematic, comprehensive, and quantitative. Gene expression profiling has revealed valuable discoveries in basic biological research, pharmacology, and medicine. Currently, clustering has become a popular method for profiling genomic data by which clusters are formed based on the similarity between data points. The points in each specific cluster are similar from each other but different from points outside this cluster.

Clustering methods depend on the measure of pair-wise similarity, the similarity between two points. One commonly used similarity metric is the correlation coefficient between the profiles of the two points, and another commonly used similarity metric is the Euclidean distance. The measure of similarity based on correlation coefficients captures the similarity in shape or pattern of the profiles, and it does not account for the amplitude of the profiles. Scaled versions of any two profiles will have the same correlation coefficient since that of the pair of original profiles, i.e., the amplitude of the profiles, does not affect the correlation coefficient as long as the wave form (shape or pattern) of the profiles is maintained. If the similarity is measured by distance, the amplitude of the profiles does matter. Two profiles with the same pattern but very different amplitudes can have an ideal similarity (essentially the same) when measured by the correlation coefficient, but a very low similarity when measured by the Euclidean distance due to the large difference in amplitudes.

In this study, we focus on clustering based on similarity measured by the correlation coefficient where two genes with similar expression patterns will be

considered to be similar regardless of the difference in their amplitudes. Using the Pearson correlation coefficient as the similarity metric, Eisen *et al.* (Eisen, Spellman et al. 1998) analyzed one of the first genome-wide microarray data sets for the budding yeast *Saccharomyces cerevisiae*. When calculating the gene expression similarity with the Pearson correlation coefficient (Eisen, Spellman et al. 1998), many studies only averaged the replicates in each experiment (Kung, Kenski et al. 2005; Rengarajan, Bloom et al. 2005; Matsumura, Bin Nasir et al. 2006) without taking into account the error in the replicates. Instead of averaging over the replicates, Hughes *et al.* (Hughes, Marton et al. 2000) defined an error model which uses a standard deviation (SD)-weighted correlation coefficient (SDCC) to down-weight the gene expression values with high error estimates in their clustering analysis and classified the functions of previously uncharacterized genes by comparing the expression profiles of mutant cells from their *S. cerevisiae* compendium. Using the same correlation, van't Veer *et al.* (van't Veer, Dai et al. 2002) derived a breast cancer prognosis from the gene expression profile of a primary tumor. In addition, Yeung *et al.* (Yeung, Medvedovic et al. 2003) showed that the SD-weighted correlation coefficient improves cluster accuracy and stability to a greater extent than the Pearson correlation coefficient with averaging replicates.

However, the SD-weighted correlation coefficient (Hughes, Marton et al. 2000; van't Veer, Dai et al. 2002; Yeung, Medvedovic et al. 2003) also has disadvantages. The error of measurement is estimated directly by the standard deviation of the replicates, and such an estimate of error can be very inaccurate when the number of replicates is small relative to the number of objects (in this study, genes) (Schäfer and Strimmer 2005). Unfortunately, most of the microarray experiments performed by an academic laboratory employ only small (usually less than 10) number of replicates due to the experimental cost and time concerns, and such a replicate number is much smaller than the amount of

genes profiled (usually in the thousands or more). The "Stein phenomenon" (Stein 1956) suggests that an effective statistical model is needed to deal with replicated microarray data. Here we provide a shrinkage correlation coefficient that considers both the number of replicates and the variance within each experimental group and fully exploits the similarity between the replicated microarray experimental samples. The shrinkage concept is widely accepted as a method to improve the estimation of correlation when the sample size is small (Stein 1956; James and Stein 1961; Ledoit and Wolf 2004), which is the primary inspiration for this work. We first describe our shrinkage correlation coefficient in generality, and then demonstrate the superiority of our correlation compared to the other two most widely-used correlation coefficients (Pearson correlation coefficient and SD-weighted correlation coefficient) using hierarchical clustering and k-means clustering. Finally we use a recently published analysis of the gene expression changes that occur during the germination of spores of the fern *Ceratopteris richardii* (Salmi, Bushart et al. 2005) as an example of this shrinkage correlation coefficient.

Other clustering techniques have been used for gene expression analysis. For example, using an analysis of variance (ANOVA) model, Kerr and Churchill (Kerr and Churchill 2001) applied bootstrapping on a publicly available data set to assess the reliability of clustering results. Ng *et al.* (Ng, McLachlan et al. 2006) proposed a linear mixed-effects model (LMM) as an extension of the normal mixture model to incorporate covariate information into the clustering process. Tjaden (Tjaden 2006) developed a clustering method that is similar to k-means clustering. Using a Bayesian infinite mixture model (IMM), Medvedovic and Sivaganesan (Medvedovic, Yeung et al. 2004) developed a clustering procedure to incorporate the information on experimental variability into gene expression profiling. IMM measures the similarity using a probabilistic model of the data, and the error between repeated measurements is inherently represented in the

model. Instead of forming final clusters directly, a posterior distribution of the possible clusters is generated by Gibbs sampling first. Then the similarity between two data points is measured by the probability of the pair of points being in the same cluster inferred by the posterior distribution of the clustering result. With this measured similarity, final clusters are formed by applying the classical hierarchical clustering. Conceptually, IMM considers the magnitude (rather than pattern) of gene expression profiles and is similar to the Euclidean distance, as noted by Tjaden (Tjaden 2006). In contrast, as a correlation coefficient, our method is based on the pattern of gene expression profiles and is apparently different from IMM when applied to clustering methods.

We stress that our shrinkage correlation coefficient is a correlation instead of a clustering method, and it could be used as a similarity metric in many clustering methods or other circumstances. Therefore, we compare our correlation with two existing widely-used correlation coefficients (Pearson correlation coefficient and SD-weighted correlation coefficient) using hierarchical and k-means clustering. We present our method for better estimating the error in replicated microarrays that cannot be adequately estimated by other correlation coefficients when applying the existing popular clustering methods.

In this chapter, we propose a novel correlation coefficient, shrinkage correlation coefficient (SCC). The comparison of SCC with other two most widely-used correlation coefficients using hierarchical and k-means clustering shows that SCC is an alternative to the Pearson correlation coefficient and the SD-weighted correlation coefficient and achieves better clustering performance on both synthetic expression data as well as real gene expression data from *Saccharomyces cerevisiae*. We use SCC-based two-dimensional hierarchical clustering to analyze the replicated microarray data of Salmi et al. (Salmi, Bushart et al. 2005), revealing the novel finding that there are two distinct clusters of genes with shared expression patterns during the early stages of germination



of *C. richardii* spores. Findings from this gene expression analysis suggest that some of the mechanisms that control germination in such diverse plant lineages as mosses and angiosperms are also conserved among ferns.

## METHODS

### Data Source, Retrieval, and Missing Value Imputation

Data analyzed here were collected from spotted cDNA microarrays produced by our lab. TUG (tentative unique gene) expression changes in *Ceratopteris richardii* were studied during the emergence from dormancy over the first 48 hr of spore germination using microarrays representing an estimated 3,207 distinct genes from this organism (Salmi, Bushart et al. 2005). Four different pairwise developmental time point comparisons were conducted with a minimum of eight replicates for each comparison: 0 vs. 24 hr, 6 vs. 24 hr, 12 vs. 24 hr, 48 vs. 24 hr. The reference sample was 24 hr for these experiments. Total RNA samples from each time point were labeled during reverse transcription with one of the fluorescent Cy5 (red) or Cy3 (green) dyes (Amersham Biosciences, Buckinghamshire, UK). Experimental design, including probe synthesis, hybridization conditions and array scanning can be found in a published protocol (Salmi, Bushart et al. 2005). Dye-swap experiments (biological replicates) were included for all time point comparisons. Raw data, array images, settings, grid files, red/green scan files, compiled tabular data, detailed protocols are publicly available from the Longhorn Array Database (LAD) (Killion, Sherlock et al. 2003).

It should be noted that for the 12:24 hr time point comparison group, two prints of arrays (Cri2 and Cri3) were used for hybridization. These arrays were the same except printed on different days. After four replicates were conducted on the Cri2 arrays, the new Cri3 arrays were used for the remaining five replicates as described (Salmi, Bushart et al. 2005).

Spots with aberrant measurements due to array artifacts or poor quality were manually flagged, and spots contaminated with dust or fluorescent specks were excluded from further analysis. The  $\log_2$  of background-subtracted, normalized median spot

intensities of ratios from the two channels (Cy5/Cy3) were retrieved from LAD (Killion, Sherlock et al. 2003) after filtering out spots that had weak signal intensities based on the following criteria: the regression correlation value between the signal intensities in the two channels (Cy5 and Cy3) across all pixels was required to be greater than 0.5, and the sum of median intensities for the two channels was required to be greater than 150. Spots that meet the above criteria had to make up at least 80% of the array for it to be included in further analysis. To focus this analysis on the TUGs with the greatest changes in expression, we selected TUGs whose fluorescence intensity ratio (in at least two replicate arrays of any time point comparison) differed by  $\geq 1.5$ -fold from their geometric mean ratio across the entire set of arrays.

Any missing values in arrays included in analysis were imputed by the K-nearest neighbors (KNN) algorithm (Troyanskaya, Cantor et al. 2001) with the average value of the nearest ten neighbors ( $K = 10$ ). The Euclidean distance was used to determine the nearest neighbors for a given gene. These imputed values were used throughout the analysis but were left blank in the primary data tables.

### **Singular Value Decomposition**

We used singular value decomposition (SVD) to uncover the artifacts in the data set that were caused by comparison of different biological replicates and prints of arrays. Let  $X$  denote an  $m \times n$  real matrix with  $m \geq n$ , the singular value decomposition of the rectangular matrix  $X$  is a factorization of the form,  $X = USV^T$ , where  $U$  is an  $m \times n$  orthogonal matrix,  $S$  is an  $n \times n$  diagonal matrix with non-negative entries in non-increasing fashion, and  $V^T$  is the transpose of  $V$ , an  $n \times n$  orthogonal matrix. The diagonal entries of  $S$  are the singular values. The columns of  $U$  and the rows of  $V^T$  are called the left- and right-singular vectors, respectively. The singular vectors form

orthonormal bases. Each left-singular vector is mapped onto the corresponding right-singular vector with the corresponding singular value. Also, each left-singular vector (or right-singular vector) is completely uncorrelated with all the other left-singular vectors (or right-singular vectors). The fraction of singular value,  $p_i = S_i^2 / \sum_{j=1}^n S_j^2$ , indicates the relative variance captured by the  $i$ th singular value (and the corresponding left- and right-singular vectors). Following the convention of (Alter, Brown et al. 2000), we refer to the left-singular vectors as eigenarrays, the right-singular vectors as eigengenes, and the singular values as eigenexpressions. All of the SVD analyses were implemented with MATLAB (MathWorks, Inc., Natick, MA).

### Shrinkage Correlation Coefficient

Correlation coefficients are computed to measure the similarity between each pair of genes in hierarchical clustering. Assuming that we have a total of  $N$  arrays consisting of  $F$  experimental (in this study, time point comparison) groups with  $N(k)$  replicates for the  $k$ th experimental group, then  $N = \sum_{k=1}^F N(k)$ . Let  $G_{i,n}(k)$  denote the expression level of the  $i$ th gene for the  $n$ th replicate in the  $k$ th experimental group. The mean and variance of the expression of the  $i$ th gene over the replicates in the  $k$ th experimental group are defined as  $\bar{G}_i(k) = \sum_{n=1}^{N(k)} G_{i,n}(k) / N(k)$  and  $S_i^2(k) = \frac{1}{N(k)-1} \sum_{n=1}^{N(k)} (G_{i,n}(k) - \bar{G}_i(k))^2$ , respectively.

If the standard deviation (SD) is used as an estimate of the measurement error, then the SD-weighted average expression of gene  $i$  over the experimental groups is given by

$$\bar{G}_i = \frac{\sum_{k=1}^F \bar{G}_i(k)}{\sum_{k=1}^F S_i^2(k)} \quad (1)$$

and the SD-weighted correlation coefficient is defined as

$$\rho_{ij}^{EW} = \frac{\sum_{k=1}^F \frac{(\bar{G}_i(k) - \bar{G}_i)}{S_i(k)} \frac{(\bar{G}_j(k) - \bar{G}_j)}{S_j(k)}}{\sqrt{\sum_{k=1}^F \left( \frac{\bar{G}_i(k) - \bar{G}_i}{S_i(k)} \right)^2 \sum_{k=1}^F \left( \frac{\bar{G}_j(k) - \bar{G}_j}{S_j(k)} \right)^2}}, \quad (2)$$

which has been used in the previous studies (Hughes, Marton et al. 2000; van't Veer, Dai et al. 2002). Since the SD-weighted correlation takes the measurement error into account, it is better than the Pearson correlation coefficient in estimating the correlation between a pair of genes in the case of repeated measurements (Hughes, Marton et al. 2000; van't Veer, Dai et al. 2002; Yeung, Medvedovic et al. 2003). The concept of applying larger weights on genes with smaller measurement error seems natural and has been demonstrated to be effective. However, the standard deviation may not be the best estimate of measurement error according to the Stein Phenomenon (James and Stein 1961; Efron and Morris 1973).

If  $F > 2$ , the Stein estimation, defined as  $\tilde{S}_i^2(k) = (1 - \frac{F-2}{D})S_i^2(k)$  with  $D = \sum_{k=1}^F (S_i^2(k))^2$ , is better than the standard variance in the sense that it is statistically closer to the real error (Efron and Morris 1973). The last statement is valid under the assumption that  $S_i^2(k)$ ,  $k=1,2,\dots,F$ , are Gaussian distributed and independent of each other, which can be readily justified using the central limit theorem.

Since the Stein Phenomenon was discovered, several shrinkage estimation methods have been developed to find the optimal estimate of a group of measurement errors. In this work we propose a simple but effective methodology which is mainly inspired by the shrinkage concept (Stein 1956; James and Stein 1961; Ledoit and Wolf 2004).

Let the real squared measurement errors for the  $F$  experimental groups be  $\psi_i(1), \psi_i(2) \dots \psi_i(F)$ . We may estimate these  $F$  parameters using a high-dimensional model (of dimension  $F$ ). According to statistics theory, the estimates in a high dimensional model will have larger variances compared to those in low-dimensional model (e.g., one dimension) when the same number of data points are available. Furthermore, if the number of data points are very limited (as is typically the case in real examples) the variances of a high-dimensional model may be unacceptably high for practical purposes. To reduce the estimation variance one may map the high-dimensional model for the  $F$  parameters onto a lower-dimensional restricted submodel. For example we may use the mean of the  $F$  parameters,

$$\Theta_i = \frac{1}{F} \sum_{k=1}^F \psi_i(k) \quad (3)$$

as a one-dimensional submodel. Then, the estimation variance can be greatly reduced. However, the estimate is biased if we replace  $\psi_i(1), \psi_i(2) \dots \psi_i(F)$  by  $\Theta_i$ .

To summarize, the estimates in the original high-dimensional model have larger variances but are unbiased, and the estimate in the restricted one-dimensional submodel has a smaller variance but is biased. Since neither situation is satisfactory, we will propose a shrinkage error estimate that makes a balance between the above two kinds of estimates.

With the above restricted one-dimensional model (Eq. 3), we have an unbiased estimate of the squared measurement error  $\Theta_i$  as follows:

$$\bar{S}_i^2 = \frac{1}{N-F} \sum_{k=1}^F \sum_{n=1}^{N(k)} (G_{i,n}(k) - \bar{G}_i(k))^2 = \sum_{k=1}^F \frac{N(k)-1}{N-F} S_i^2(k). \quad (4)$$

Then, we can use a linear regularization model to define a balanced estimate:

$$T_i(k) = (1 - \lambda_i) S_i^2(k) + \lambda_i \bar{S}_i^2, \quad (5)$$

where  $\lambda_i \in [0, 1]$  is an shrinkage factor that is to be determined according to a chosen optimization criterion. We propose to minimize the risk

$$R(\lambda_i) = E\left(\sum_{k=1}^F (T_i(k) - \psi_i(k))^2\right). \quad (6)$$

Applying the methodology of (Ledoit and Wolf 2004) and (Schäfer and Strimmer 2005), the optimal shrinkage factor can be derived as

$$\hat{\lambda}_i = \frac{\sum_{k=1}^F [\text{var}(S_i^2(k)) - \text{cov}(S_i^2(k), \bar{S}_i^2)]}{\sum_{k=1}^F E[(S_i^2(k) - \bar{S}_i^2)^2]}. \quad (7)$$

From previous discussion,  $S_i^2(k)$ ,  $k=1,2,\dots,F$ , are assumed to be independent of each other, and from the above equation we have

$$\hat{\lambda}_i = \frac{\sum_{k=1}^F [\text{var}(S_i^2(k)) - \text{cov}(S_i^2(k), \bar{S}_i^2)]}{\sum_{k=1}^F (S_i^2(k) - \bar{S}_i^2)^2} = \frac{\sum_{k=1}^F (1 - \frac{N(k)-1}{N-F}) \text{var}(S_i^2(k))}{\sum_{k=1}^F (S_i^2(k) - \bar{S}_i^2)^2}, \quad (8)$$

where  $\text{var}(S_i^2(k))$  is the variance of  $S_i^2(k)$  estimated by

$$\text{var}(S_i^2(k)) = \frac{N(k)}{(N(k)-1)^3} \sum_{n=1}^{N(k)} [(G_{i,n}(k) - \bar{G}_i(k))^2 - S_i^2(k)]^2. \quad (9)$$

To make sure that the shrinkage factor lies between 0 and 1, we define the final shrinkage factor to be

$$\lambda_i^* = \min(1, \max(0, \hat{\lambda}_i)), \quad (10)$$

and then we obtain the shrinkage estimate of  $S_i^2(k)$ ,  $k=1,2,\dots,F$ , as

$$T_i^*(k) = (1 + \lambda_i^* S_i^2(k))^{-1} S_i^2(k). \quad (11)$$

Notice that in (Schäfer and Strimmer 2005), a related mathematical problem is considered, where it aims to get a shrinkage estimate of a covariance matrix by using a restricted lower dimensional submodel in which the covariance matrix is assumed to be diagonal with common variance. In this work our problem is simpler. We only need to estimate the diagonal elements of the covariance matrix, not the whole matrix. Therefore, our result is different from what is obtained in (Schäfer and Strimmer 2005), and is much simpler.

Using  $T_i^*(k)$ , the shrinkage estimate of  $S_i^2(k)$ , the error between the group mean  $\bar{G}_i(k)$  and the corresponding true expression value can be measured by means of the shrinkage error defined as  $\Phi_i(k) = \sqrt{\frac{T_i^*(k)}{N(k)}}$ . If we replace  $S_i(k)$  by  $\Phi_i(k)$  in Eqs. 1 and 2, the shrinkage error-weighted average expression of gene  $i$  is given by

$$\bar{G}_i^{sw} = \sum_{k=1}^F \frac{\bar{G}_i(k)}{\Phi_i^2(k)} \bigg/ \sum_{k=1}^F \frac{1}{\Phi_i^2(k)}, \quad (12)$$

and we can define a new shrinkage correlation coefficient for any pair of  $i$ th and  $j$ th genes as

$$\rho_{ij} = \frac{\sum_{k=1}^F \frac{(\bar{G}_i(k) - \bar{G}_i^{sw})(\bar{G}_j(k) - \bar{G}_j^{sw})}{\Phi_i(k)\Phi_j(k)}}{\sqrt{\sum_{k=1}^F \left( \frac{\bar{G}_i(k) - \bar{G}_i^{sw}}{\Phi_i(k)} \right)^2 \sum_{k=1}^F \left( \frac{\bar{G}_j(k) - \bar{G}_j^{sw}}{\Phi_j(k)} \right)^2}}. \quad (13)$$

If the number of replicates  $N(k)$  is the same for all the experimental (e.g., treatment/condition/time-point) groups, then  $\Phi_i(k) = \alpha T_i^*(k)$ , where  $\alpha = 1/N(k)$  is a constant, and hence the shrinkage correlation coefficient (Eq. 13) is effectively weighted by the shrinkage estimate  $T_i^*(k)$  (i.e.,  $N(k)$  has no effect on the shrinkage correlation



coefficient). However, if the number of replicates is different for each experimental group, the use of  $\Phi_i(k) = \sqrt{\frac{T_i^*(k)}{N(k)}}$  in the shrinkage correlation coefficient provides additional benefits in our method through the weighting  $N(k)$  in a way that agrees with the common practice in statistics (Bland 1995).

When using a biologically meaningful control sample, e.g., a matched and untreated sample, a zero time point, or the reference sample 24 hr presented in this study, we should use uncentered correlation to keep the impact of the biologically meaningful control sample on the gene expression changes (Demeter, Beauheim et al. 2007). Therefore, we replace  $\bar{G}_i^{sw}$  and  $\bar{G}_j^{sw}$  with zero in our analysis, and hence Eq. 13 is modified as follows:

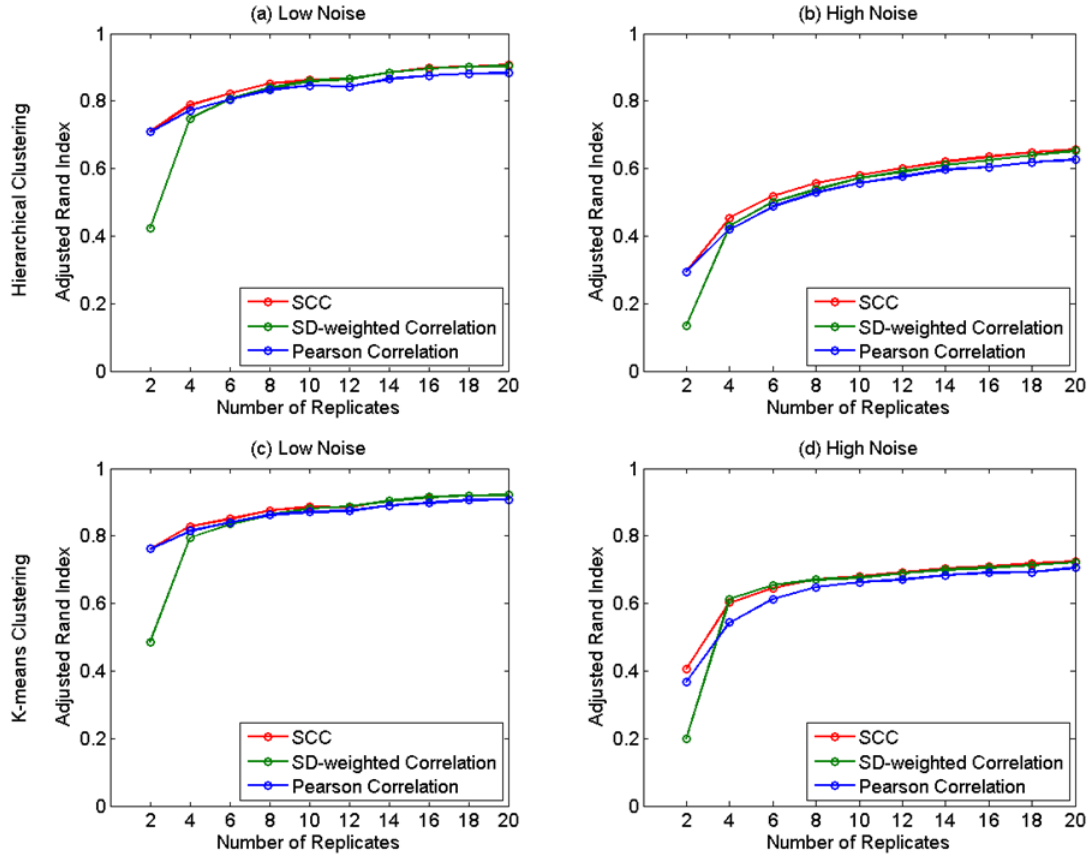
$$\rho_{ij} = \frac{\sum_{k=1}^F \frac{\bar{G}_i(k)}{\Phi_i(k)} \frac{\bar{G}_j(k)}{\Phi_j(k)}}{\sqrt{\sum_{k=1}^F \left( \frac{\bar{G}_i(k)}{\Phi_i(k)} \right)^2 \sum_{k=1}^F \left( \frac{\bar{G}_j(k)}{\Phi_j(k)} \right)^2}}. \quad (14)$$

**Evaluation of SCC.** As shown in Eq. 13, SCC is a type of correlation coefficient which is very useful in gene expression clustering. To assess the effectiveness of a new correlation coefficient in clustering analysis, it is important to compare it with other widely used correlation coefficients using existing popular clustering methods. In this study, we compared SCC with the two most commonly used correlation coefficients: Pearson correlation coefficient (Eisen, Spellman et al. 1998) and SD-weighted correlation coefficient (Hughes, Marton et al. 2000; van't Veer, Dai et al. 2002; Yeung, Medvedovic et al. 2003). We applied these three correlation coefficients on the two most popular clustering methods: hierarchical clustering (Hartigan 1975) and k-means clustering (MacQueen 1967), and evaluated the performance by comparing the adjusted Rand index

(Hubert and Arabie 1985) generated for each correlation using these two clustering methods. Both synthetic expression data and real yeast expression data (Ideker, Thorsson et al. 2001) were used in this study. The adjusted Rand index is a statistic that has been recently used for the comparison of clustering using different correlation coefficients (Yeung, Haynor et al. 2001; McShane, Radmacher et al. 2002; Kasturi, Acharya et al. 2003; Monti, Savage et al. 2005; Ng, McLachlan et al. 2006). It measures the extent of concurrence between the clustering results and the underlying known cluster structure (Milligan and Cooper 1986). The comparison of the adjusted Rand indices generated by different correlation coefficients for the same data set indicates the performance of each correlation. The adjusted Rand index lies between 0 and 1, and a larger index indicates a higher level of agreement between the clustering results and the prior knowledge of functional categories, and further suggests better clustering performance (Yeung, Medvedovic et al. 2003; Monti, Savage et al. 2005).

Each synthetic data set includes 20 experiments, 2, 4, 6, 8, 10, 12, 14, 16, 18, and 20 replicates for each experiment, and two different levels of noises (low and high). The data sets were generated with predetermined patterns plus low or high level of random noise so that the underlying cluster structure is known. We evaluated the level of agreement between the resulting clusters from each of the three correlations and the known underlying cluster structure by computing the average adjusted Rand index over 1000 randomly generated synthetic data sets. Figure 2.1a and b show the correlation comparisons using hierarchical clustering and Figure 2.1c and d are the results from k-means clustering. As shown in Figure 2.1a, under low noise level ( $\alpha = 0.5$  in Eqs. 15 and 16), SCC has the same performance as the Pearson correlation coefficient but far better than the SD-weighted correlation coefficient when the replicate number is two. When the number of replicates is four, six and eight, SCC is superior to the other

correlations. With the increasing of the number of replicates, the SD-weighted correlation coefficient approaches SCC but both of them still outperform the Pearson correlation coefficient. Figure 1b shows that, when the noise level is high ( $\alpha = 2.5$  in Eqs. 15 and 16), SCC performs the best for all the numbers of replicates. These results suggest that, for synthetic microarray data, SCC is superior to the Pearson correlation coefficient and the SD-weighted correlation coefficient when using hierarchical clustering as it results in the most consistently high adjusted Rand index regardless of noise level and number of replicates.

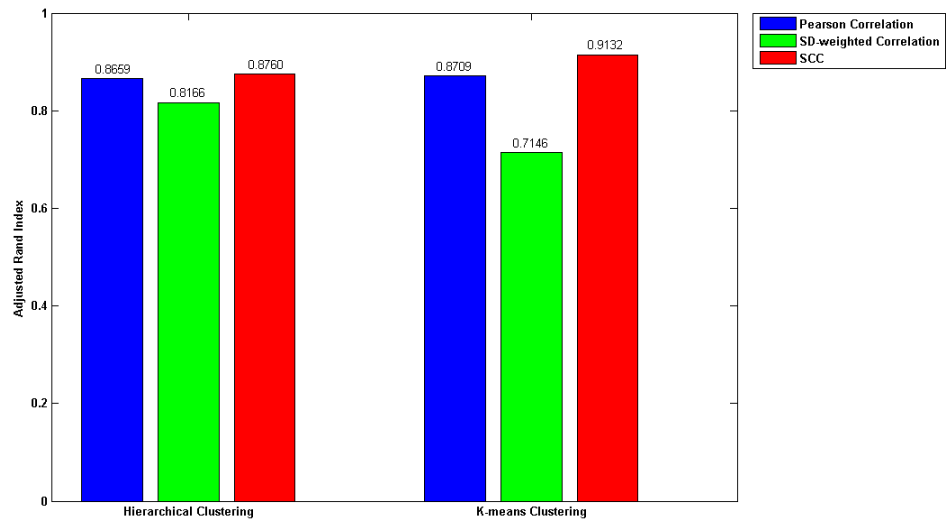


**Figure 2.1.** The performance of the three models indicated by the adjusted Rand index obtained from the synthetic data sets using hierarchical clustering and k-means clustering. The number of the replicates varies from 2 to 20. Each correlation is represented by a curve: SCC (red), SD-weighted correlation (green), and Pearson correlation (blue). Every data point on a curve is an average adjusted Rand index over 1000 trials of generating and clustering the synthetic data. Hierarchical clustering: (a) Low noise level. (b) High noise level. K-means clustering: (c) Low noise level. (d) High noise level. Error bars are not shown here because, given the scaling of the Figure, they are too small to be graphically depicted after 1000 trials.

We also compared the correlations using k-means clustering. Under low noise level (Figure 2.1c), SCC surpasses the other two correlations when the number of replicates is lower than 12. With the increasing of the number of replicates, the SD-weighted correlation coefficient approaches SCC but both of them still outperform the Pearson correlation coefficient. While the noise level is high (Figure 2.1d), SCC outperforms the Pearson correlation coefficients for almost all the numbers of replicates. When compared with the SD-weighted correlation coefficient, SCC is better when the number of replicates is smaller than four and close to the SD-weighted correlation coefficient when the number of replicates is four and six. With the increasing of the number of replicates, we noticed that SCC performs almost equally as the SD-weighted correlation coefficient and has a slight advantage when the number of replicates is larger than 14. The k-means clustering results suggests that, SCC is a better choice compared to other correlations when the expression noise level is low, while the noise level is high, SCC is obviously superior to the Pearson correlation coefficient on almost all the numbers of replicates and has slight advantages over the SD-weighted correlation coefficient for most of the numbers of replicates.

To further demonstrate the superiority of SCC, we applied the three correlations individually to hierarchical and k-means clustering and computed the adjusted Rand index on the real yeast expression data (Ideker, Thorsson et al. 2001). This microarray data represent 20 systematic perturbations of the yeast galactose-utilization pathway, and four replicates were performed for each perturbation. Each gene has been annotated in one of four functional clusters in the Gene Ontology (Ashburner, Ball et al. 2000). These four clusters are used as the external knowledge. As shown in Figure 2.2, when hierarchical clustering is used, the adjusted Rand indices for SCC is 0.8760 which is higher than those of the other two correlations (Pearson correlation coefficient: 0.8659;

SD-weighted correlation coefficient: 0.8166). When applied to k-means clustering, SCC is also superior to other correlations with the highest adjusted Rand index 0.9132. Since the noise level of this real yeast expression data was not clearly stated and barely quantified in the previous study, we could not determine which is better between the Pearson correlation coefficient and the SD-weighted correlation coefficient. However, this real expression data comparison suggests that SCC is superior regardless of noise level and clustering methods which corresponds to the results with the synthetic data.



**Figure 2.2.** The performance of the three correlations indicated by the adjusted Rand index obtained from the real yeast expression data using hierarchical clustering and k-means clustering. Each correlation is represented by a bar: SCC (red), SD-weighted correlation (green), and Pearson correlation (blue). The y-axis is the adjusted Rand index.

## Gene Ontology/Functional Enrichment Analysis

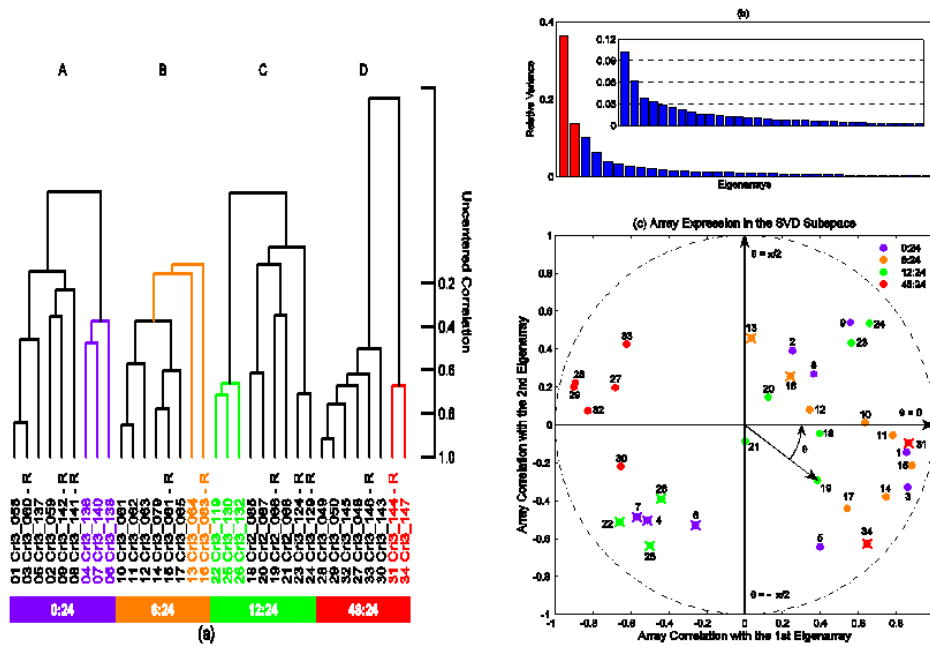
Over-representation analysis of Gene Ontology annotations associated with clusters was performed using the “ORA” analysis option in version 2.12 of the ErmineJ software (Lee, Braynen et al. 2005). Briefly, this analysis uses a re-sampling approach to compute empirical p values for each GO annotation associated with cluster members, followed by multiple hypothesis testing correction using the Benjamini-Hochberg false discovery rate (FDR) adjustment (Benjamini and Hochberg 1995). Terms with FDR less than or equal to 0.1 were considered as significantly enriched. ErmineJ requires a microarray annotations file that relates array identifiers (Genbank and TUG ids) to Gene Ontology codes and a GO term definition file (gene\_ontology.obo), available from the Gene Ontology Web site. Note that ErmineJ observes the “True Path” rule of the Gene Ontology in that annotation with a child GO term implies annotation by all its ancestor terms (Ashburner, Ball et al. 2000). Thus, parental terms that are not explicitly cited in the microarray annotations file may be found to be significantly-enriched. To create the microarray annotations file, we performed a provisional GO annotation of the *C. richardii* cDNAs (Genbank ids) using results from a prior blastx analysis in which the *C. richardii* sequences were searched against an *Arabidopsis thaliana* protein sequence database. GO terms associated with the putative *A. thaliana* homologs identified by the blastx analysis were transferred to the *C. richardii* clones. GO annotations for *A. thaliana* were obtained from the Gene Ontology Web site in November, 2006. For the GO over-representation analysis, cluster members were compared with the full set of TUGs represented on the array.



## RESULTS

### ***C. richardii* Microarray Data Quality Control**

We first applied unsupervised agglomerative hierarchical array clustering to measure the relative similarity among the replicates across all the arrays for each developmental time point comparison. The hierarchical clustering analyses were carried out simultaneously but separately on the four different time point comparison groups represented by the entire 34 arrays. Figure 2.3a shows four distinct dendrograms derived from unsupervised hierarchical clustering analysis using the traditional similarity metric, uncentered correlation (D'Haeseleer 2005) and average linkage (Eisen, Spellman et al. 1998). The biological replicates in each of the four pairwise time point comparison groups were sorted according to the degree of similarity in expression patterns across all TUGs by one-dimensional array clustering.



**Figure 2.3.** Identification of low-quality arrays. (a) Unsupervised hierarchical clustering of the 572 filtered TUGs across 34 arrays. The dendrograms describe the degree of relatedness between arrays, with shorter branches denoting a higher degree of similarity. Under the dendrograms, the horizontal colored boxes delimit four pairwise time point comparison groups. A) 0:24 hr (*violet box*). B) 6:24 hr (*orange box*). C) 12:24 hr (*green box*). D) 48:24 hr (*red box*). Sample names and branches of the outlining samples representing experimental artifacts for each of the five dendrograms are similarly color coded. For each individual group, the samples with name “- R” denote dye-swap replicates. The scale to the right of the dendrograms depicts the uncentered correlation coefficient represented by the length of the dendrograms branches connecting pairs of nodes. (b) Relative variance captured by each eigenarray in the SVD-reduced “Eigenarrays” space. The relative variance captured by the first two significant

eigenarrays are depicted by red bars, and the remainders by blue bars. (c) Array expression in the two-dimensional SVD subspace. Array correlation with the first eigenarray ( $x$ -axis) vs. that with the second eigenarray ( $y$ -axis). The dashed circle outlines normalized array expression in the two-dimensional SVD subspace. The total 34 colored dots denote all of the arrays from the four pairwise time point comparison groups: 0:24 hr (*violet*), 6:24 hr (*orange*), 12:24 hr (*green*), 48:24 hr (*red*). The ten crossed dots denote the samples that might be correlated with systematic biases within the original 34 arrays.

In dendrogram A, all the three reverse replicates (dye-swapped samples) were clustered together with the 1<sup>st</sup>, 2<sup>nd</sup> and 5<sup>th</sup> replicates, while the remaining three replicates formed a distinct small cluster. Since the dye-swapped samples were used both as a source of replication and as a control of dye specific bias, grouping the three reverse replicates together with the other three replicates indicates these six samples are most similar to each other and are sufficient for further analysis of the 0:24 hr time point comparison group. The 4<sup>th</sup>, 6<sup>th</sup> and 7<sup>th</sup> replicates could be excluded from the further analysis. Dendrogram B shows that while six of the eight 6:24 hr replicates formed a group, including one dye-swapped sample, two arrays were not included in this group. These two arrays, represented by red branches, have very low mean correlation coefficients ( $\sim 0.11$  and  $\sim 0.14$ , respectively) with others. These low correlations suggest that these corresponding samples should not be included in the further analysis. In dendrogram C, 12:24 hr array replicates were split into two distinct subgroups, one including all the three dye-swapped samples and the 18<sup>th</sup>, 20<sup>th</sup>, and 21<sup>st</sup> replicates, the other composed of only three regular replicates. Grouping the three dye-swapped samples together with three regular replicates indicates these six samples share sufficient similarities for further analysis of the 12:24 hr time point comparison. 48:24 hr replicates were compared in dendrogram D. One dye-swapped sample and one other replicate clustered on a branch distinct from all other samples. The mean correlation coefficients between each of these two samples and the rest are -0.52 and -0.41, respectively, which strongly suggest there are systematic biases existing during the corresponding hybridizations. As a whole, we kept replicates that have higher degrees of similarity from each of the four groups. The uncovered ten samples that have less similarities with other samples might be associated with systematic biases and should be removed from the original data set for further analysis.

To provide further support for the finding of ten samples that might be associated with systematic biases, and to explore the relationships among replicates in each time point comparison group, we performed SVD on the original 34 arrays. SVD is an important factorization of a rectangular matrix and can linearly reduce the input data set of high dimensionality to a lower-dimensional space, which still captures a large fraction of the variance present in the original data (Golub and Van Loan 1996). It has seen wide use in the analysis of gene expression data and has proved useful in linear modeling of gene expression (Alter, Brown et al. 2000; Holter, Mitra et al. 2000), cell sample and gene classification (Nielsen, West et al. 2002), gene network modeling (Yeung, Tegner et al. 2002), experimental artifact uncovering (Nielsen, West et al. 2002; Klevecz, Bolen et al. 2004). By projecting the expression arrays to the most significant eigenarrays, SVD can classify the arrays into different groups based on their correlations with these eigenarrays. Arrays with similar expression patterns but with different amplitudes can appear to cluster more tightly. SVD has proven to be a useful method for classifying expression arrays into different cell cycle stages of *S. cerevisiae* (Alter, Brown et al. 2000; Holter, Mitra et al. 2000). Here, we reduced the original “*TUGs* × *Arrays*” space to the “*Eigenarrays*” space which spans the space of the array expression profiles. In the SVD-reduced “*Eigenarrays*” space, we calculated the relative variance captured by each eigenarray. As shown in Figure 2.3b, the first and second eigenarrays accounted for 50% of the total variance of the data, and might be used to approximate the gene expression changes observed throughout the early stages of gametophyte development of *C. richardii*.

We further investigated the array expression profiles projected to the normalized two-dimensional SVD subspace that is spanned by the first two eigenarrays (Figure 2.3c). The ten samples identified above that might be associated with systematic biases (Figure

2.3a) are noticeably separated from the rest of the samples in each time point comparison group. For example, for the 48:24 hr time point comparison group, the arrays separated into two distinct groups: one included 6 arrays which were all anticorrelated with the first and most significant eigenarray, and the other 2 arrays (31<sup>st</sup> and 34<sup>th</sup> arrays in Figure 2.3a) represented by two crossed red dots were both correlated with the first eigenarray and anticorrelated with the second eigenarray. We infer that these 31<sup>st</sup> and 34<sup>th</sup> arrays represent the experimental artifacts caused by the hybridizations. This finding corroborates the hierarchical array clustering result shown in dendrogram D in Figure 2.3a. In the 0:24 hr and 12:24 hr time point comparison groups, all the dubious arrays outlined previously in Figure 2.1a are clustered together in the third quadrant, while the remaining arrays from those two time point comparison groups spread in the first and fourth quadrants. For the 6:24 hr time point comparison group, although all of the arrays are in the first and fourth quadrants, the 13<sup>th</sup> and 16<sup>th</sup> arrays outlined previously in Figure 2.3a tend to be away from the others. This independent assessment made by SVD is consistent with the above finding by one-dimensional hierarchical array clustering, further indicating those ten samples might be correlated with systematic biases in the hybridizations and most likely represent experimental artifacts.

Furthermore, the array expression profiles projected to the normalized two-dimensional SVD subspace show that arrays from the first three time point comparison groups (0:24 hr, 6:24 hr, and 12:24 hr) express in the first and fourth quadrants in the entire two-dimensional SVD subspace, while the arrays from the 48:24 hr group express in the second and third quadrants. Their opposite correlations with the most significant eigenarray suggest that the first three time point comparison groups may have similar expression profiles which are opposite with that of the 48:24 hr time point comparison group.

### **Cluster Analysis of *C. richardii* Gene Expression**

Spores of *C. richardii* have proved to be an excellent model system to study the basic cellular processes that occur in early gametophyte development, such as gravity sensing and response, sex determination and differentiation, pattern formation, and photomorphogenesis (Chatterjee and Roux 2000). Using DNA microarrays consisting of 3,840 spotted cDNA clones from an EST analysis, Salmi *et al.* (Salmi, Bushart et al. 2005) monitored the mRNA levels for 3,207 tentative unique genes (TUGs) of *C. richardii* over the first 48 hr of gametophyte development. TUG expression in the spores was evaluated at 0, 6, 12, 24 and 48 hr after spores were exposed to continuous white light. This developmental period includes initiation of germination at 0 hr, the production of a detectable polar calcium current that peaks at 6-12 hr, fixation of the polarity of more than half of cells by gravity at 12 hr, migration of the nucleus at 24 hr, and a polar cell division at 48 hr (Chatterjee, Porterfield et al. 2000).

The data analysis of Salmi *et al.* (Salmi, Bushart et al. 2005) focused on identification of differentially-expressed genes between time points and did not attempt to identify upward or downward trends in the time course data. This analysis did not discover and filter out the underlying biases associated with experimental artifacts due to comparison of different biological replicates and prints of arrays to facilitate further gene expression profiling. Moreover, this analysis did not organize expression patterns into biologically meaningful profiles through the whole time course experiments by assimilating the patterns of gene expression. A more thorough analysis of these microarray expression data that focuses on characterizing the entire set of transcripts temporally and displaying them graphically would promote a better understanding of cellular processes underlying early gametophyte development in ferns.

In total, we analyzed 34 arrays with biological replicates of four different developmental time point comparisons: nine replicates of 0:24 hr, eight of 6:24 hr, nine of 12:24 hr, and eight of 48:24 hr. The 34 arrays were used to generate a total of 34 data columns in which each column was treated independently rather than averaging the replicates.

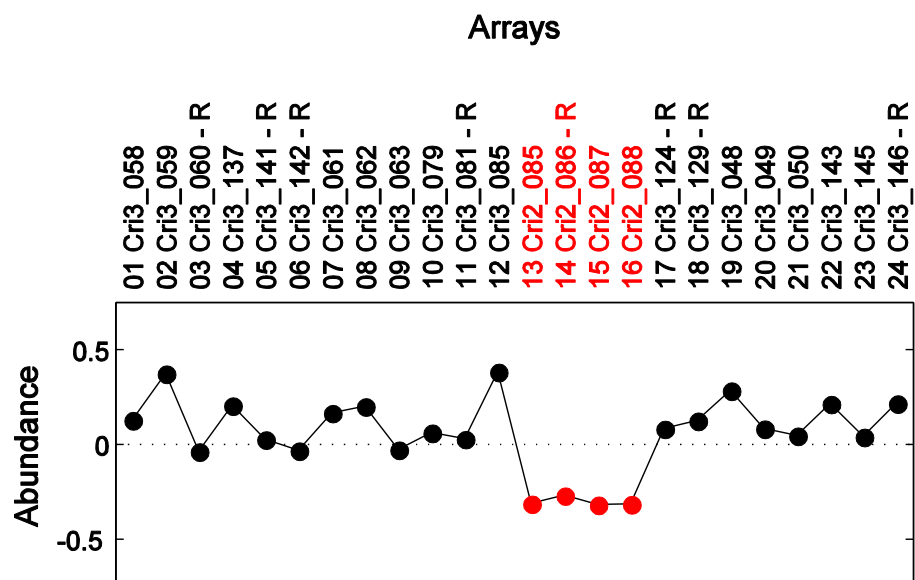
The initial selection retained non-flagged spots for which the within-spot pixel-to-pixel correlation of intensities is  $> 0.5$  and the sum of median (635/532) signal intensities is  $> 150$ . These non-flagged spots were also well-measured in at least 80% of the array. In this selection, 39% of the TUGs were filtered out prior to further analysis. Selecting for TUGs with a known accession number in GenBank removed a further 4% of the TUGs. Selection for a fluorescence ratio of at least 1.5-fold greater than the geometric mean ratio for the TUGs examined in at least two arrays of any time point comparison removed 41%. The resulting data set (see Additional file 1) for the experiments analyzed included tabulation of the  $\log_2$  ratios of gene-expression levels for 572 TUGs. In the entire 34 arrays there were 137 TUGs for which there were no missing data. A total of 1,152 expression ratios were missed, accounting for only about 6% of the total data set we analyzed. After the imputation of the missing data with the K-nearest neighbors (KNN) method (Troyanskaya, Cantor et al. 2001), the data set was normalized array-wise with the mean of 572 TUGs expression ratios of each array set as zero and the standard deviation as one.

We first uncovered ten samples that have poor correlation to other samples of their time point group by applying unsupervised agglomerative hierarchical array clustering and SVD. These ten arrays, which likely represent experimental artifacts, were removed from the data set, and the new data set was used to tabulate the  $\log_2$  ratios of gene-expression levels for 572 TUGs. Of these TUGs, 151 have no missing data in the



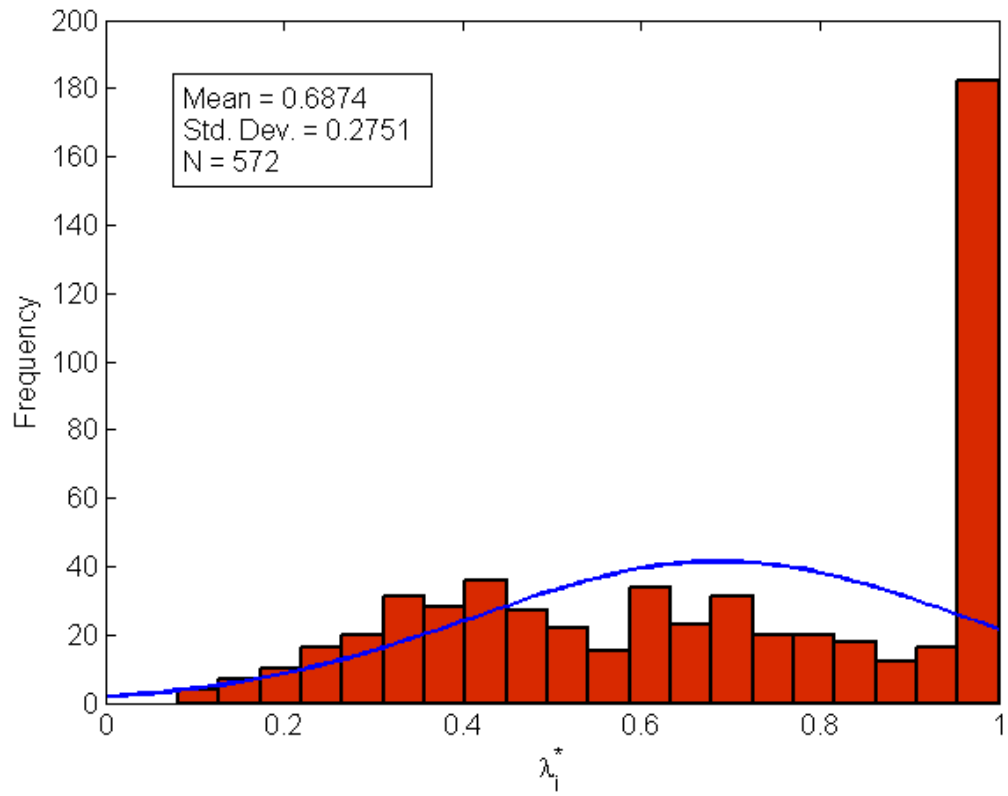
entire 24 arrays. A total of 1020 data were missed, accounting for only 7.4% of the adjusted data set. The imputation of the missing data was performed by the K-nearest neighbors (KNN) method (Troyanskaya, Cantor et al. 2001) with the average value of the nearest 10 neighbors ( $K = 10$ ). We further normalized this new data set by adjusting the mean of 572 TUGs expression ratios of each array to zero and the standard deviation to one.

In the new data set, the expression profile of the 12:24 hr time point comparison group was measured by two prints of arrays: four replicates hybridized on the Cri2 arrays, and two on the Cri3 arrays. Cri2 and Cri3 arrays were the same except printed on different days. In attempt to identify and correct the gene-wise bias introduced by the two prints of arrays, we carried out SVD. This technique has previously been used to detect and correct the artifact in the data set that was caused by different types of arrays (22K vs. 42K) (Nielsen, West et al. 2002) or sampling from cultures with slightly different periods (Li and Klevecz 2006). We reduced the new “*TUGs* × *Arrays*” space to the “*Eigenarrays*” space that spans the space of the array expression profiles and the “*Eigengenes*” space that spans the space of the gene expression profiles. Eigengene 5 was discovered to be exactly correlated with the gene-wise bias (Figure 2.4). We sorted the abundance of this eigengene in each of the 24 arrays, and found a perfect correlation between the abundance and the print of array (Figure 2.4). All of the four Cri2 arrays show negative abundances, whereas the Cri3 arrays all have non-negative abundances.

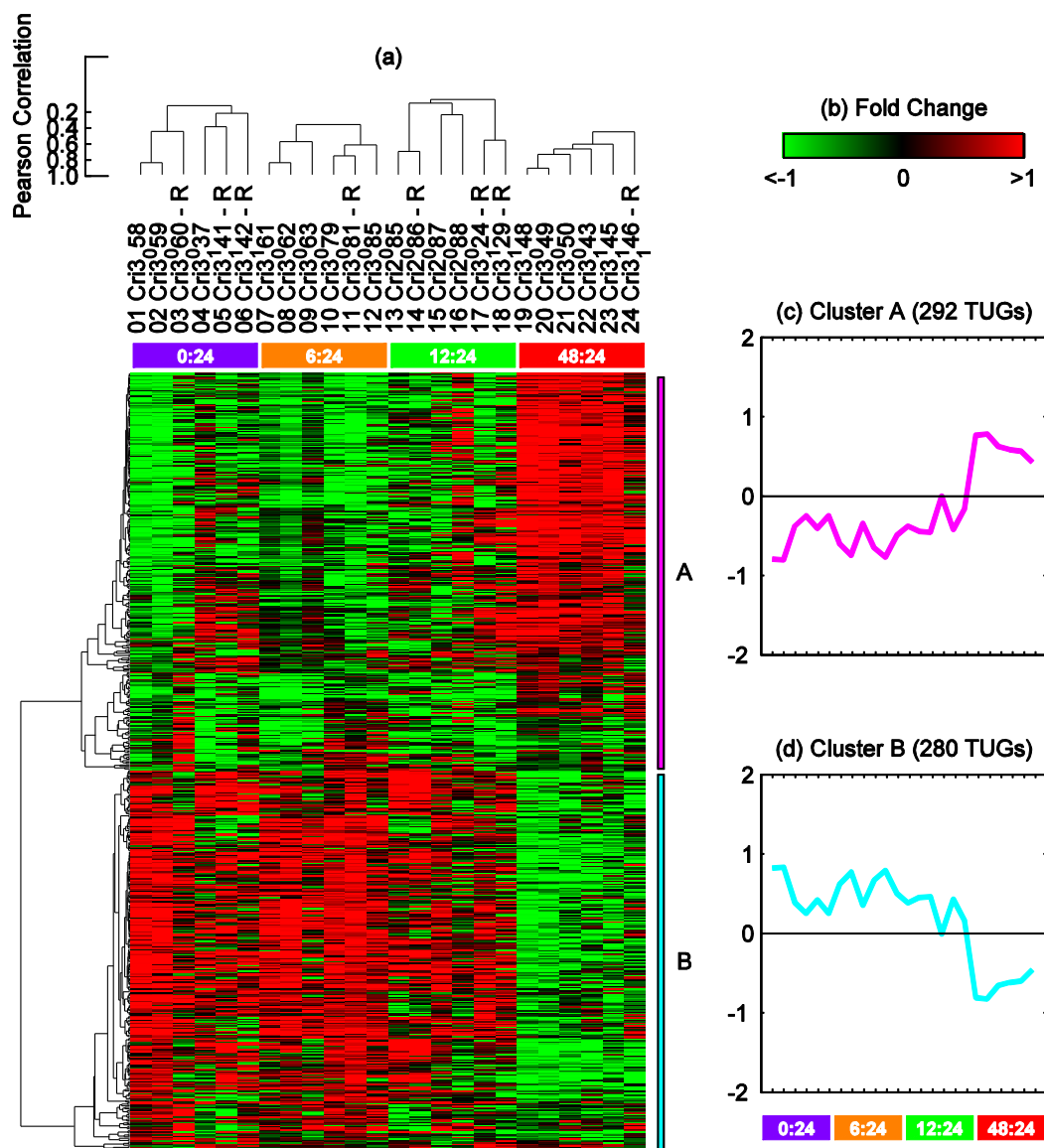


**Figure 2.4.** Gene-wise bias (Eigengene 5) associated with the two prints of arrays. The abundance of Eigengene 5 in each of the 24 arrays with the arrays in the order obtained in Additional file 3. The 24 dots denote all of the arrays: Cri2 arrays (*red*), Cri3 arrays (*black*). Array names are similarly color coded.

We filtered out the gene-wise bias from the data set by substituting Eigenexpression 5 with zero, and reconstructed the final data set. Subsequently, this final data set was analyzed by the unsupervised two-dimensional agglomerative hierarchical clustering. We first calculated the optimal shrinkage factor  $\lambda_i^*$  with Eq. 10. Since the number of replicates is six (a moderate number) for each time point comparison after we filtered out ten low-quality arrays,  $\lambda_i^*$  is significantly different from 0 and 1, and has an average value of 0.69 (Figure 2.5). This indicates SCC is effective in our analysis, and neither the Pearson correlation coefficient nor the SD-weighted correlation coefficient should be used here. Therefore, we applied SCC to the gene clustering, and the Pearson correlation coefficient (Eisen, Spellman et al. 1998) to the array clustering. The TUG expression profiles during the early stages of gametophyte development of *C. richardii* are shown in Figure 2.6. The 572 TUGs are clearly clustered into two distinct groups: Cluster A with 292 TUGs, and Cluster B with 280 TUGs.



**Figure 2.5.** Histogram of optimal shrinkage factor  $\lambda_i^*$ . The mean, standard deviation, and the total number of  $\lambda_i^*$  are shown in the left upper corner of the histogram.



**Figure 2.6.** TUG expression profile in the early stages of gametophyte development of *C. richardii* by SCC. (a) Unsupervised two-dimensional hierarchical clustering. Data are presented in a matrix format: each row represents an individual TUG, and each column corresponding to an experimental sample. Each expression measurement represents the normalized  $\log_2$  ratio of fluorescence from the hybridized experimental sample to a reference sample. Normalized TUG expression ratios are depicted by a pseudocolor scale

with red indicating positive expression above the reference, black indicating equal expression as the reference, and green indicating negative expression below the reference. The horizontal colored boxes delimit four pairwise time point comparison groups: 0:24 hr (*violet box*), 6:24 hr (*orange box*), 12:24 hr (*green box*), and 48:24 hr (*red box*). The scale to the left of the dendrograms depicts the Pearson correlation coefficient represented by the length of the dendrograms branches connecting pairs of nodes. (b) The fold change scale extends from fluorescence ratios of -1 to 1 in  $\log_2$  units. (c) Average expression profiles of Cluster A, computed by averaging the  $\log_2(\text{Cy5/Cy3})$  ratios. (d) Average expression profiles of Cluster B, computed by averaging the  $\log_2(\text{Cy5/Cy3})$  ratios.

We performed functional analysis of clusters A and B using Gene Ontology annotations transferred from putative *A. thaliana* homologs of individual TUGs, as identified by blastx analysis described previously (Salmi, Bushart et al. 2005). Using a standard over-representation analysis as implemented in the ErmineJ software (Lee, Braynen et al. 2005), we examined the clusters to identify Gene Ontology terms that appeared unusually often among the TUGs in each cluster, relative to the full complement of TUGs represented on the array.

We found that Cluster A, which includes genes that are upregulated during the first 48 hours following germination, were significantly enriched with genes annotated with the GO term RNA-binding. The cluster contained TUGs that had high homology to *A. thaliana* proteins At4g32720.1 (CriU1545, CriU226) and At2g05120.1 (CriU2095). At4g32720.1 contains an RNA recognition motif, and At2g05120.1 contains a region matching a nucleoporin Pfam motif. Both are annotated with the term “RNA export from nucleus,” suggesting that their *C. richardii* counterparts may also be involved in RNA processing and export.

Over-representation analysis of Cluster B, which includes genes that are downregulated during the same period of development, revealed fourteen enriched terms. These included terms related to signal transduction (protein phosphatase type 2C activity, abscisic-acid mediated signaling, hormone-mediated signaling), transport, biosynthetic activities (water-soluble vitamin biosynthesis), oxidative phosphorylation, and response to oxidative stress. A full list of terms is given in Table 2.1.

GO id	Term	Genes	p value (ORA)
<b>Cluster A</b>			
GO:0050658	RNA transport	5	0.0002
<b>Cluster B</b>			
GO:0009615	response to virus	8	0.0001
GO:0008047	enzyme activator activity	5	0.0002
GO:0030312	external encapsulating structure	5	0.0002
GO:0017077	oxidative phosphorylation	5	0.0002
	uncoupler activity		
GO:0008287	protein serine/threonine phosphatase complex	9	0.0003
GO:0015071	protein phosphatase type 2C activity	6	0.005
GO:0016311	dephosphorylation	11	0.001
GO:0042221	response to chemical stimulus	48	0.0011
GO:0005730	nucleolus	8	0.0022
GO:0042364	water-soluble vitamin biosynthesis	8	0.0022
GO:0009738	abscisic acid mediated signaling	8	0.0022
GO:0006810	transport	92	0.0022
GO:0005886	plasma membrane	25	0.0027
GO:0009755	hormone-mediated signaling	14	0.0035

**Table 2.1. GO categories significantly (FDR < 0.10) enriched among genes belonging to SCC Clusters A and B. GO id:** the Gene Ontology identifier for each over-represented Gene Ontology terms. **Term:** the name of the term. **Genes:** Number of genes on the array with the GO annotation term in column 1. **P value:** The unadjusted p value from the over-representation analysis (ORA) test.



In addition, we noticed that the final data set analyzed here was generated by filtering out ten low-quality microarray samples and one gene-wise bias from the original 34 arrays. Therefore, it will be reasonable to obtain similar clusters based on this final data set by applying the three correlations (SCC, Pearson correlation coefficient and SD-weighted correlation coefficient) to hierarchical or k-means clustering. Since SCC is superior in clustering synthetic and real expression data as demonstrated above, we presented here the application of SCC to analyze our own microarray data.

## DISCUSSION

### **Shrinkage Correlation Coefficient is a Robust Correlation**

Correlation coefficient is crucial in cluster analysis to determine the similarity between two objects (in this study, genes) and further classify the objects into different groups. When the Pearson correlation coefficient was first applied for clustering gene expression (Eisen, Spellman et al. 1998), the replicates of each treatment group were simply averaged without considering the underlying error. As the importance of the error information was discovered (Hughes, Marton et al. 2000), more and more studies use standard deviation as the error estimate when clustering gene expression with the help of correlation coefficient. Using the SD-weighted correlation coefficient, they down-weighted the gene expression values with high error estimates in microarray analysis (Hughes, Marton et al. 2000; van't Veer, Dai et al. 2002; Yeung, Medvedovic et al. 2003). However, the SD-weighted correlation coefficient is still not statistically efficient for analyzing replicated microarray data and the use of standard deviation as the error estimate exhibits serious defects when the number of replicates is small (Schäfer and Strimmer 2005).

Commonly, the number of microarray replicates that are performed by most academic laboratories is usually less than 10 due to the experimental cost and time concern. The “Stein phenomenon” (Stein 1956) states that when the number of data samples (in this study, biological replicates) in each experimental group is relatively small, a better estimate of the error of any individual experimental group could be obtained by shrinkage that considers all experimental groups. To avoid inaccuracy introduced by the small number of replicates, a better estimate of the error in the replicates can be obtained by shrinkage estimate. Our shrinkage correlation coefficient (Eq. 13) can be regarded as a generalized definition of correlation coefficient for the

expression of a pair of genes with replicates. It takes into consideration both the number of replicates and the variance within each treatment comparison.

SCC can be reduced to other definitions of correlation coefficients when the shrinkage factor  $\lambda_i^*$  is set to some special values. For example, under the condition that the numbers of replicates are equal for all the  $F$  experimental groups, SCC is equivalent to the Pearson correlation coefficient when  $\lambda_i^* = 1$ , and to the SD-weighted correlation coefficient when  $\lambda_i^* = 0$ . Therefore, the SD-weighted correlation coefficient and the Pearson correlation coefficient are just two extreme cases of SCC. The correlation coefficient with optimal shrinkage is an alternative to these two extremes and is superior to them when  $0 < \lambda_i^* < 1$ . By using the shrinkage factor  $\lambda_i^*$ , we obtain an optimal estimate of the error in the replicates and, accordingly, better estimates of the similarity between any pair of genes.

We would argue that in SCC, the use of the shrinkage error  $\Phi_i(k) = \sqrt{\frac{T_i^*(k)}{N(k)}}$  as a weighting provides a better correlation measure for two reasons. First,  $T_i^*(k)$  is superior to the standard deviation as an estimate of the measurement error, and secondly, the inclusion of  $N(k)$  takes the size of an experimental group into account as an assessment of the reliability of the group mean. The benefit of  $N(k)$  disappears if all experimental groups are of the same size. For example, in our *C. richardii* microarray data analysis, the number of replicates happens to be equal for each time point comparison after filtering out ten low-quality arrays. In such an analysis, the use of  $\Phi_i(k) = \sqrt{\frac{T_i^*(k)}{N(k)}}$  and the use of  $\Phi_i(k) = \sqrt{T_i^*(k)}$  are equivalent, since either choice leads to the same value of shrinkage correlation in Eq. 13. Figure 2.1a, b and c shows that SCC is superior to the other models regardless of the number of replicates. In Figure 2.1d, SCC also has advantages for most

of the numbers of replicates. Therefore, SCC offers utility to most replicated microarray data sets.

Using the Stein shrinkage concept (Stein 1956; James and Stein 1961), a shrinkage estimator for gene-specific variance components was proposed to construct a *F*-like statistic that has been used in a linear mixed ANOVA (analysis of variance) model (Cui, Hwang et al. 2005), and a shrinkage estimator of the mean used in the clustering similarity metric was developed for genome-wide expression data analysis (Cherepinsky, Feng et al. 2003). This shrinkage ANOVA model and the shrinkage mean could be combined with our SCC for analyzing expression data in future studies.

## **Results of Functional Analysis**

The stringent quality control filtering followed by novel cluster analysis methodology of microarray data on *C. richardii* early development produced two distinct clusters of TUGs. As shown in Figure 2.5, the TUGs represented in SCC Cluster A increased expression during the first 48 hours following germination, while TUGs in Cluster B decreased expression during the same period. An analysis of GO terms associated with TUGs in Cluster A reveals that only one term (RNA transport) is over-represented among GO annotations associated with Cluster A TUGs relative to the other TUGs assayed in the microarrays. This suggests that the TUGs in Cluster A represent a cross-section of many different types of genes, perhaps reflecting a general "ramp-up" in multiple biological functions, along with an increase in RNA processing as the spore activates embryonic transcription.

Previous results from other systems suggests that genes associated with the term RNA transport may be related to the process of establishing and maintaining cellular polarity in the early stages of germination of *C. richardii* spores. In the filamentous

fungus *Aspergillus nidulans*, mutation of *swoK*, a gene that encodes a protein with an N-terminal RNA recognition motif that causes cells to swell and lack the normal polarity maintained fungal hyphal cells (Shaw and Upadhyay 2005). The *swoK* protein appears to function in both mRNA maturation and nuclear export of mRNAs.

By contrast, over-representation analysis of the Cluster B TUGs, which are downregulated during the 48 hr time course, reveals an enrichment of several GO terms, suggesting that the Cluster B TUGs represent a more specialized set of genes involved in functions and processes required during early development of *C. richardii*. Cluster B consists of genes that are downregulated during the earliest stages of spore germination, starting with the initiation of germination by light (0 hr) through the first two days of development, when the first cell division occurs. These are likely to include transcripts that were present in the dormant spore but decline in abundance in the germinating spore. This population is likely to encode proteins involved in maintaining the dormant condition. Once germination begins, genes responsible for maintaining the dormant condition of the spore would need to be downregulated to allow for the transition from dormant metabolism to active growth and development.

Careful examination of the Cluster B genes and their associated GO terms reveals some interesting patterns. One of the most notable findings from this study is that genes involved in abscisic acid mediated signaling are overrepresented among genes downregulated in the first 48 hr of spore germination. Abscisic acid (ABA) is a plant hormone known to be involved in the process of establishing and maintaining dormancy in angiosperm seeds (Bove, Lucas et al. 2005; Kermode 2005) and moss spores (Decker, Frank et al. 2006). ABA has been previously shown to be involved in another aspect of *C. richardii* development. Sex determination of *C. richardii* gametophytes is regulated in part by ABA (Banks, Hickok et al. 1993).

The process of seed germination in *Arabidopsis* involves a decrease in the endogenous levels of abscisic acid (Kermode 2005), and inhibiting ABA biosynthesis by treatment with fluridone caused *Nicotiana plumbaginifolia* seeds that should be physiologically dormant (D) to germinate at the same rate as seeds that are in a physiologically non-dormant (ND) condition (Bove, Lucas et al. 2005). The more general term hormone-mediated signaling (a sub-set of which would be abscisic acid-mediated signaling) is also over-represented among the TUGs in Cluster B. This term annotates TUGs predicted to be involved in ABA-related pathways as well as other hormone-related processes, notably signaling pathways mediated by gibberellic acid. Hormone-mediated regulation of the process of germination has been well studied in angiosperm seeds, including ABA involvement in the maintenance of dormancy, and the germination activating role of gibberellin. The process of germination in angiosperm seeds involves extensive hormone-mediated signaling (Ogawa, Hanada et al. 2003) so it is not surprising to find this category of genes implicated in fern spore germination, as well.

The involvement of ABA in germination is not unique to angiosperm seeds. In moss cells that have differentiated into spores, removal of ABA causes these cells to germinate and develop into new filamentous cells (Schnepf and Reinhard 1997). In the classical view of hormone signaling in eukaryotic organisms hormones are thought of as the chemical substances produced in one part of the organism that serves as a signal to another part of the organism. In this paradigm it would seem unusual for a plant hormone to function within the single cell of the *C. richardii* spore. However, an ABA receptor involved in seed germination has been characterized that is not plasma membrane localized (Shen, Wang et al. 2006), the *Arabidopsis* gene CHLH (genomic locus At5g13630). This receptor functions inside of cells and it could, in principle, respond to intracellular changes in the level of ABA to regulate physiological changes within single

cells. The CHLH "receptor" gene encodes a subunit of the Mg-chelatase complex that is an integral part of chlorophyll biosynthesis, and is involved in plastid-to-nucleus signaling. The process of producing a highly resistant, dormant stage of plant reproductive cycles (i.e. a spore or seed) is conserved among all major plant lineages. Given the intercellular localization and functioning of an ABA receptor that mediates germination in *Arabidopsis* seeds, and the documented role of ABA in the germination of various spores, it is plausible that this signaling pathway is similar in fern and moss germination.

One of the well-documented mechanisms that regulates ABA signaling in seed germination is dephosphorylation of regulatory proteins by protein phosphatases, particularly type 2C protein phosphatases (PP2C) (Reyes, Rodriguez et al. 2006; Saez, Robert et al. 2006). In *Fagus sylvatica* the expression of a PP2C is directly regulated by ABA in dormant seeds (Lorenzo, Nicolas et al. 2002). Additionally, dephosphorylation of actin by protein phosphatases has been implicated in the germination of the plasmodium *Physarum sclerotium* (Furuhashi 2002) and *Dictyostelium discoideum* (Kishi, Mahadeo et al. 2000). The biological process of dephosphorylation was included as an activity that is over-represented in the cluster of genes down regulated during germination, and this category includes genes likely to encode PP2C enzymes. The prediction from our clustering results that ABA-regulated signaling enzymes like protein phosphatases are involved in fern spore germination is plausible in light of the conservation of this pathway in other plant species.

Annotation of genes in the down-regulated cluster identified three cellular localization categories, including external encapsulating structure, nucleolus, and plasma membrane. Of these three categories, the down regulation of nucleolar associated genes is similar to a process observed in angiosperm seed germination. In *Zea mays* seeds,

nucleolus-associated bodies are present in the cells of dry seeds, but after 24 h of imbibition these nuclear bodies have decreased significantly (Gulemetova, Chamberland et al. 1998). Although *C. richardii* spores are a useful model system for the study of gravity perception in a single cell, we did not identify any genes in this analysis of early development that are obviously involved in the process of gravity perception or in early signaling steps of the gravity response.



## Chapter 3 Comparative Analysis of *Arabidopsis thaliana* Microarray Data

### BACKGROUND

Ectoapyrase (ecto-NTPDase) are enzymes that remove the terminal phosphate from extracellular nucleoside triphosphates (e.g., ATP) and nucleoside diphosphates. Rapidly growing tissues in *Arabidopsis*, such as pollen tubes and etiolated hypocotyls, release ATP into their ECM as they grow, and they strongly express two nearly identical ectoapyrase proteins, APY1 and APY2, which function together to limit the concentration of extracellular ATP. RNAi-induced suppression of *APY1* gene expression in *apy2* mutants disrupts auxin transport and severely suppresses growth in *Arabidopsis* (Wu, Steinebrunner et al. 2007). To better understand the implications of these findings, an analysis of the underlying gene expression changes that accompany *APY* gene suppression was carried out. We used an inducible RNAi construct to suppress *APY1* in plants homozygous for the *apy2* knockout mutation. Growth inhibition of the mutant seedlings becomes evident after 3 days of growth in the presence of the estradiol inducer. We compared gene expression differences between uninduced plants and plants grown continuously in the inducer for 3.5 days (dark-grown) or 6 days (light-grown) using the NimbleGen *Arabidopsis thaliana* 4-Plex microarray. We compared the two sets of large-scale expression data and identified genes whose expression significantly changed after ectoapyrase suppression in light- and dark-grown plants, respectively. Major changes in numerous transcription factors and in hormone-regulated genes were observed, and five of them were independently verified by qRT-PCR. Data analysis has provided a better understanding of the molecular bases underlying the relationship of ectoapyrase expression to growth.

## **METHODS**

### **Plant Materials and Growth Conditions**

*Arabidopsis thaliana* ecotype Wassilewskija (WS) was used as wild type in this study. Seeds were surface sterilized and planted in a Murashige and Skoog medium (4.3 g/L Murashige and Skoog salts, 0.5% MES, 1% sucrose, and 1% agar, pH 5.7 with 5M KOH) (Sigma-Aldrich, St. Louis). The APY1 and APY2 mutants were isolated previously (Steinebrunner, Wu et al. 2003). For light-grown seedlings, plates were placed upright in a culture chamber and grown at 23°C under 24 hours fluorescent light for six days. And for dark-grown seedlings, plates were grown in darkness for 3.5 days.

### **Total RNA Isolation and NimbleGen Microarray Experiments**

Three biological replicates of RNA were prepared from each of the following samples: estradiol-induced and apyrase-directed RNAi construct in *apy2* plants that were wild type for APY1 but homozygous for the *apy2* knockout mutation, non-induced RNAi construct in *apy2* plants that were wild type for APY1 but homozygous for the *apy2* knockout mutation, estradiol-treated wild type seedlings, non-treated wild type seedlings. Total RNA was isolated from the 6-day-old light-grown seedlings and 3.5-day-old dark-grown seedlings by using Tri-reagent (Ambion, Austin, TX) in accordance with the manufacturer's protocol. Total RNA was sent to the NimbleGen *Arabidopsis thaliana* 4-Plex array platform (Catalog no: A4511001-00-01). Sample labeling, array hybridization, scanning, data extraction, and preliminary data analysis are performed at NimbleGen.

### **False Discovery Rate**

When a statistical hypothesis test is performed on each of thousands of genes represented in a genome-wide study, a measure of significance is calculated to test if this

gene is differentially expressed. Some of the earliest genome-wide studies used the  $p$ -value to measure significance. The  $p$ -value is the estimated probability of rejecting the null hypothesis ( $H_0$ ) of a study question when the null hypothesis is true, or the probability associated with a false positive (a gene that is declared to be differentially expressed although it is not). We define a significant result when the resulting  $p$ -value is less than our threshold, normally 0.05. Here, the threshold is also called the significance level ( $\alpha$ ) which is an arbitrary pre-specified probability. A significance level of 0.05 means that there is a 5% chance that we make the wrong decision or the result is false positive. While the significance level of 0.05 is acceptable for one test, if many tests are performed simultaneously on the data, then this  $\alpha$  can result in a large number of false positives. For example, if we apply a  $t$ -test to each of 10,000 genes in a genome-wide study, then we would expect to get 500 false positives by chance alone. This is known as the multiple testing problem.

To correct the multiple testing problem (i.e., reduce false positives), we can assign an adjusted  $p$ -value to each test, or similarly, reduce the significance level to each test. As discussed in Chapter 1, Bonferroni (Bonferroni 1935) and Sidak (Sidak 1967) corrections are two traditional methods to this problem. However, these two methods are too conservative in the sense that while they reduce the number of false positives, they also reduce the number of true discoveries (i.e., significant results). The False Discovery Rate (FDR) (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001; Storey and Tibshirani 2003) is a recent development that can control the number of true discoveries, but at the same time, reduce the number of false positives. The  $q$ -value is the name given to the adjusted  $p$ -value using a FDR method. The  $q$ -value is a measure of significance in terms of the FDR, whereas the  $p$ -value is a measure in terms of the false positive rate. If we choose 0.05 as a threshold, a  $p$ -value of 0.05 implies that 5% of all tests will result in

false positives, whereas a  $q$ -value (i.e., FDR adjusted  $p$ -value) of 0.05 implies that 5% of significant tests will result in false positives. The latter is clearly a far smaller quantity and tells more about the content of the so-called significant genes.

### **Significance Analysis of Microarrays**

Significance Analysis of Microarrays (SAM) (Tusher, Tibshirani et al. 2001) is a statistical technique for finding statistically significant genes in a set of microarray experiments. It considers the relative change of each gene expression level with respect to the standard deviation of repeated measurements and then assigns a score to each gene. This analysis uses a non-parametric test and a permutation idea to estimate the percentage of genes identified by chance (the false discovery rate). It does not require the assumptions of equal variances or independence of genes. The algorithm can be stated as follows:

- “1: Fix a threshold for differentially expressed genes
- 2: Count the number of differentially expressed genes in each permutation
- 3: Calculate the median number of false positives across all permutations
- 4: Calculate the FDR as the number of false positives divided by the number of genes in the original data.” (Draghici 2003)

SAM is run as an Excel Add-In and available for download online at <http://www-stat.stanford.edu/~tibs/SAM/>.

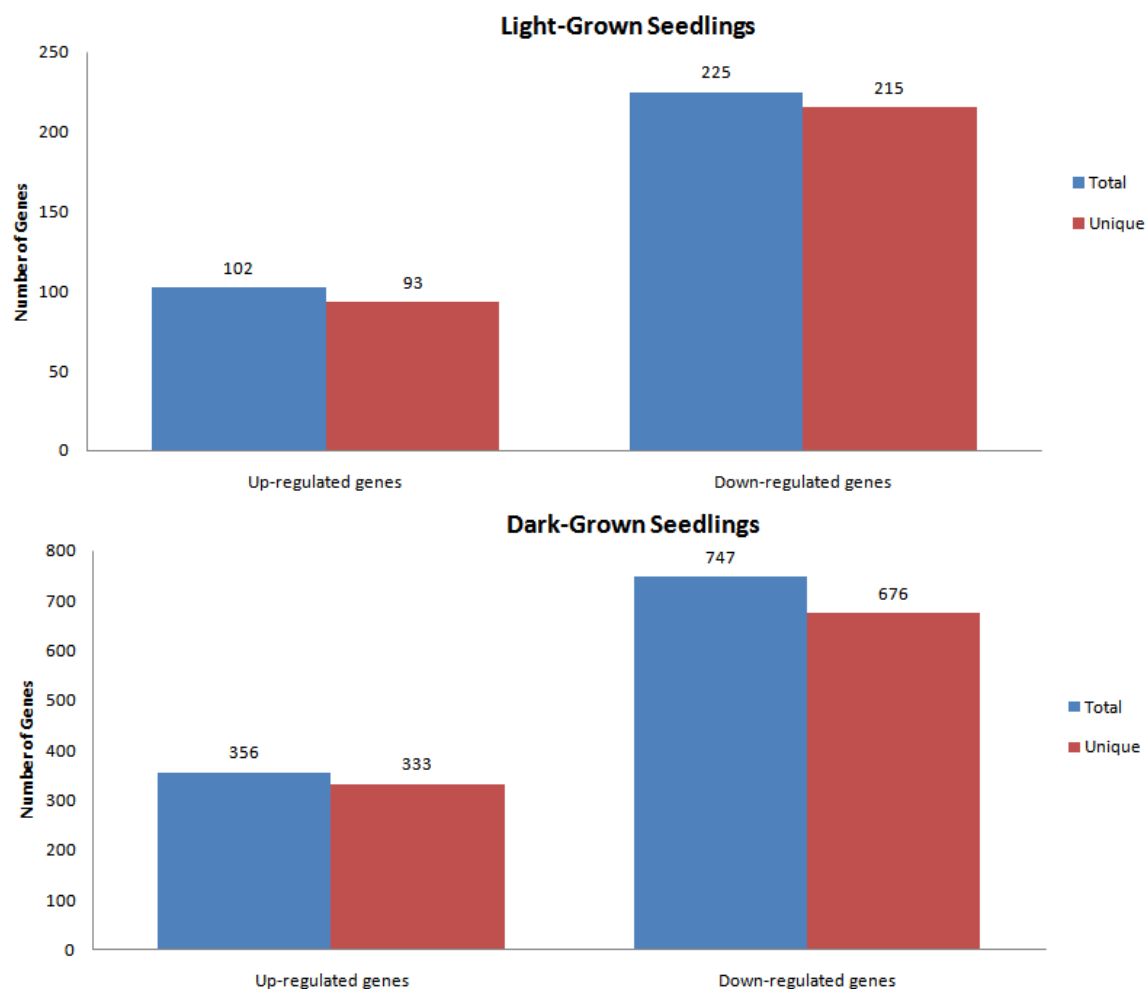
## RESULTS

To investigate *APY1* and *APY2* gene function in whole plants, it is important to characterize the phenotype of double-knockout (DKO) plants. However, DKO progeny could not be produced easily because DKO pollen cannot germinate (Steinebrunner, Wu et al. 2003). Therefore, as a complementation strategy, we generated an apyrase-directed RNAi construct in *apy2* plants that were wild type for *APY1* but homozygous for the *apy2* knockout mutation. We referred to this sample as a DKO in which *APY1* was silenced and *APY2* was knocked out. We also generated a non-induced RNAi construct in *apy2* plants that were wild type for *APY1* but homozygous for the *apy2* knockout mutation. In this SKO (single-knockout) sample, *APY2* was knocked out while *APY1* was normal. For control, we chose estradiol-treated wild type seedlings (WT+E) and non-treated wild type seedlings (WT).

To identify early response genes and define the potential roles of *APY1* and *APY2* in regulating plant growth, we performed microarray analysis of the gene expression changes that occur when the expression of apyrases is genetically suppressed and growth is severely inhibited. The hypocotyls of dark-grown plants elongate much faster than they do in light-grown plants. When the RNAi construct is induced by estradiol, the inhibition of growth becomes apparent in dark grown plants by 3.5 days, but in the slower-growing light grown plants this inhibition only becomes apparent by 6 days. Apyrase expression is higher in dark-grown hypocotyls than in light-grown hypocotyls, so after RNAi-induced suppression of apyrase, the decrease in apyrase expression is much greater in dark-grown plants than in light grown plants, thus the growth effects occur less quickly and are less dramatic in light-grown plants. It is also possible that the products of photosynthesis in light-grown plants delay the effects of apyrase suppression of growth in these plants. Since we were interested in identifying

genes involved in mediating the suppression of growth that occurs after decreased apyrase expression, we assayed for these genes after 3.5 days in dark-grown plants, but after 6 days in light-grown plants.

Two sets of NimbleGen microarray data were collected from light- and dark-grown seedlings, respectively. Three comparisons were performed in each set of microarray data to identify differentially expressed genes caused by DKO, SKO, and estradiol treatments, separately: DKO vs. WT+E, SKO vs. WT, and WT+E vs. WT. SAM (Tusher, Tibshirani et al. 2001) was applied to select differentially expressed genes. Differentially expressed genes are defined as those that have FDR q-values less than 5% and a minimum two-fold change. As shown in Figure 3.1.A, among 102 up-regulated genes in light-grown seedlings, 93 of them were solely upregulated by the DKO treatment and 9 were upregulated by the SKO or Estradiol treatment. And, among 225 down-regulated genes, 215 were solely downregulated by the DKO treatment. In dark-grown seedlings, as shown in Figure 3.2.B, 333 genes were solely upregulated by the DKO treatment and accounted for 93.5% of the total number of genes upregulated by the same treatment. Also, 676 genes were solely downregulated by the DKO treatment in dark-grown seedlings and counted for 90.5% of the total number of genes downregulated by the same treatment. Since we are primarily interested in the effect of suppressing both AY1 and APY2 on plant growth, we will pay more attention on genes solely expressed by the DKO treatment in this work.

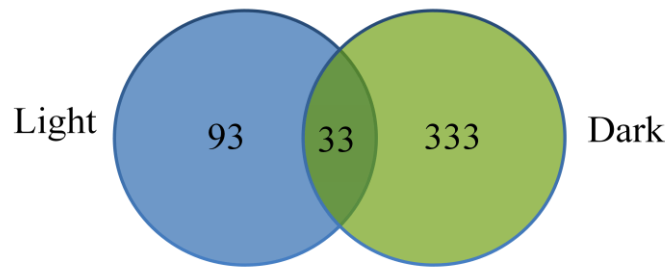


**Figure 3.1.** Number of genes altered by DKO treatment in light-grown and dark-grown seedlings, respectively. The total number of genes up- or down-regulated in each type of seedlings is represented by a blue bar. The number of genes that uniquely change expression in response to DKO treatment (i.e., those that do not change expression after SKO or Estradiol treatment) is represented by a red bar.

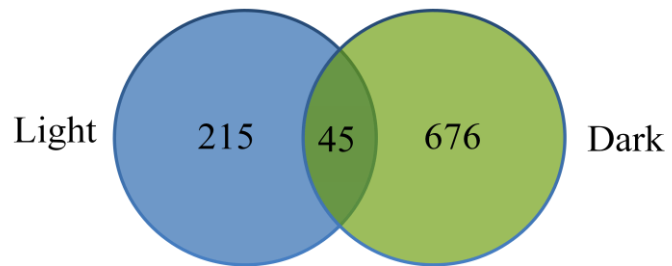
We compared the genes that were up- and down-regulated as a result of the DKO treatment in light- and dark-grown seedlings, respectively. As shown in Figure 3.2, among all up-regulated genes, 33 of them were upregulated both in light- and dark-grown seedlings. Still in Figure 3.2, 45 genes were downregulated in both seedlings. We then categorized these genes by performing Gene Ontology analysis. Figure 3.3.A shows that, among 33 up-regulated genes expressed in both types of seedlings, 10.3% of them have gene products located in plasma membrane and 7.8% are located in cell wall. As shown in Figure 3.3.B, 10.3% of the 33 up-regulated genes are related to transcription, and 5.2% of them have responses to stress and 3.5% of them have responses to abiotic or biotic stimulus. Figure 3.3.C shows that, 10.2% of the 33 up-regulated genes are related to transcription factor activity. Similarly, in Figure 3.4, we can notice that 5.9% of the 44 down-regulated genes expressed in both types of seedlings are located in nucleus, 21.7% are related to response to stress or abiotic or biotic stimulus, and 10.2% of them are involved in transcription factor activity. Full lists of the 33 up-regulated and 45 down-regulated genes are given in Tables 3.1 and 3.2.



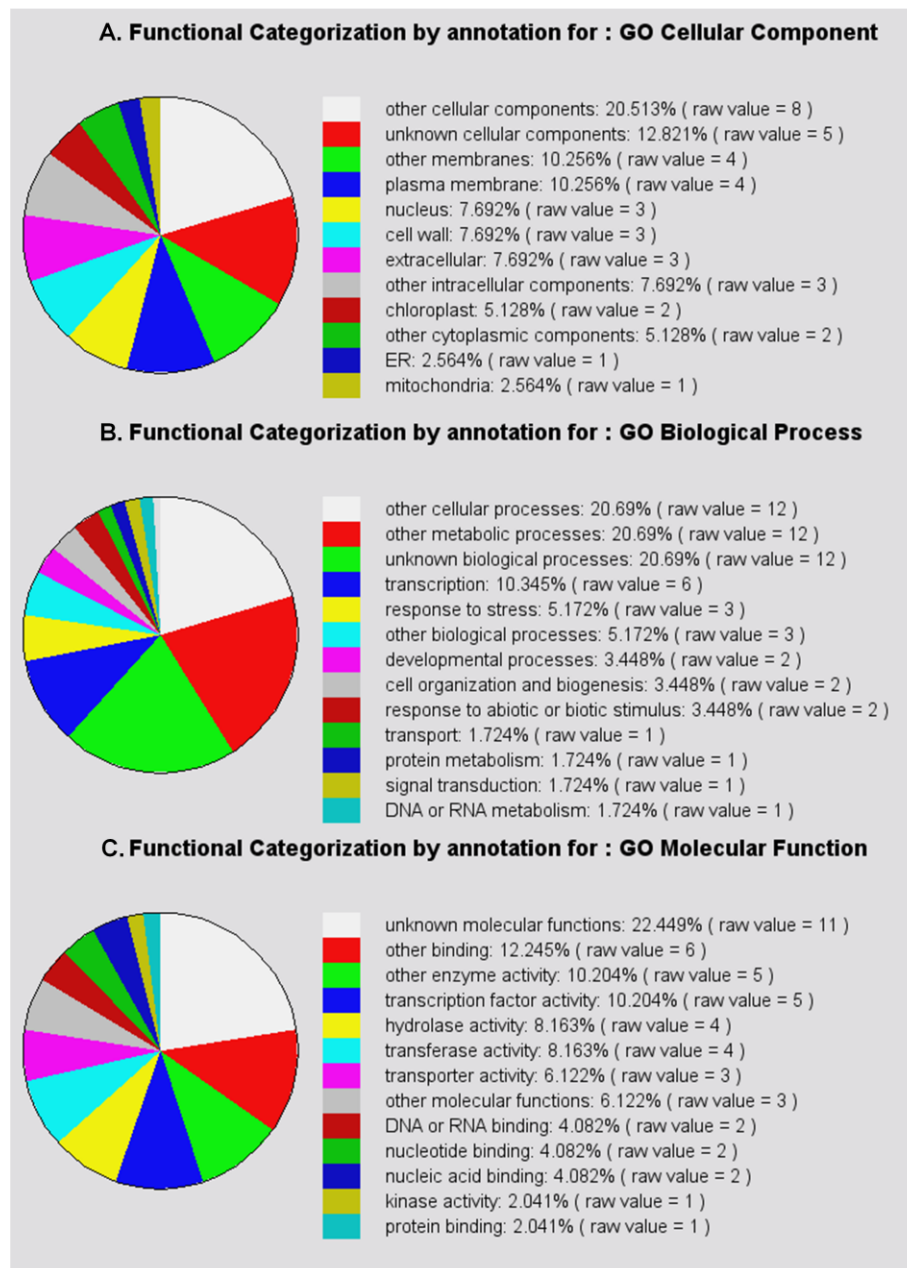
### Number of Up-regulated Genes



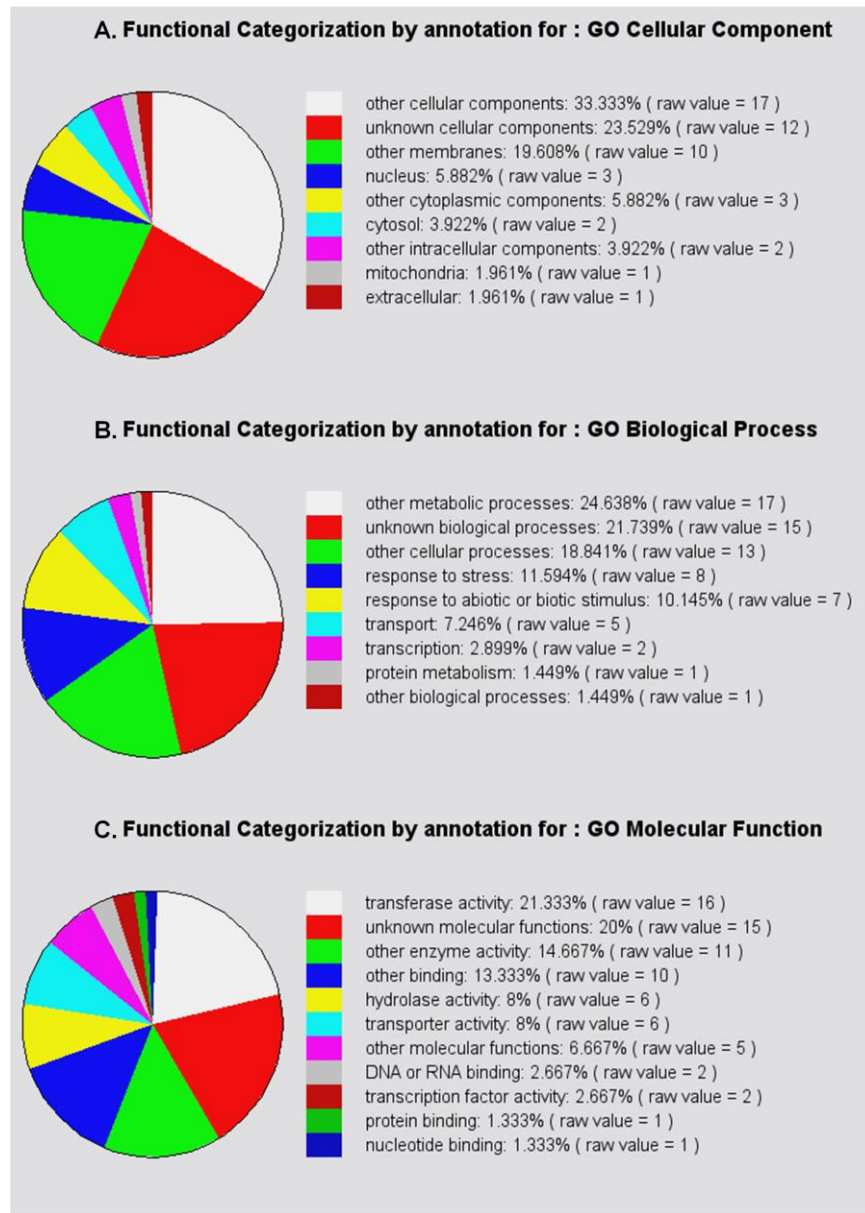
### Number of Down-regulated Genes



**Figure 3.2.** The number of up- or down-regulated genes uniquely altered by DKO treatment in light- and dark-grown seedlings.



**Figure 3.3.** Functional categorization of the up-regulated 33 genes expressed both in light- and dark-grown seedlings by the DKO treatment.



**Figure 3.4.** Functional categorization of the down-regulated 45 genes expressed both in light- and dark-grown seedlings by the DKO treatment.

Overlap Genes	Description	FDR q-value (%)	
		Light	Dark
AT4G22960	unknown protein	0	0
AT5G64060	Arabidopsis NAC domain containing protein 103; transcription factor.	0	0
AT5G61740	member of ATH subfamily	0	0.16
AT2G38340	encodes member of DREB subfamily A-2 of ERF/AP2 transcription factor family.	0	0
AT5G15380	Encodes methyltransferase involved in the de novo DNA methylation and maintenance of asymmetric methylation of DNA sequences.	0.90	0
AT4G33720	pathogenesis-related protein, putative.	0.90	0
AT5G13210	unknown protein	0.90	0
AT5G60250	zinc finger (C3HC4-type RING finger) family protein.	0.90	0
AT1G06520	Encodes mitochondrial localized protein with glycerol-3-phosphate acyltransferase activity.	0.90	0
AT3G47480	calcium-binding EF hand family protein.	0.90	0
AT4G36430	peroxidase, putative; Identical to Peroxidase 49 precursor (PER49)	0.90	0
AT5G06720	peroxidase, putative; Identical to Peroxidase 53 precursor (PER53)	0.90	0
AT4G39830	L-ascorbate oxidase, putative.	0.90	0
AT5G26300	meprin and TRAF homology domain-containing protein with MATH domain	1.37	0
AT1G20180	unknown protein	1.85	0
AT1G21520	unknown protein	2.37	0
AT4G05380	AAA-type ATPase family protein.	2.37	0
AT2G18150	peroxidase, putative; Identical to Peroxidase 15 precursor (PER15).	3.29	0
AT5G46730	glycine-rich protein.	3.29	0.86
AT2G46400	member of WRKY Transcription Factor.	3.29	0.20
AT1G21250	cell wall-associated kinase, may function as an ECM signaling receptor	3.29	0
AT1G60030	xanthine/uracil permease; Identical to Nucleobase-ascorbate transporter 7.	3.29	0
AT3G12410	3'-5' exonuclease/ nucleic acid binding.	3.29	0
AT1G49860	Encodes glutathione transferase belonging to the phi class of GSTs.	3.29	0
AT2G16450	F-box family protein; Identical to F-box protein At2g16450.	3.32	0
AT5G53230	unknown protein	3.32	0
AT3G01600	ANAC044 (Arabidopsis NAC domain containing protein 44); transcription factor.	3.91	0
AT5G55490	Encodes transmembrane domain containing protein that is expressed in pollen.	3.91	0
AT3G58270	meprin and TRAF homology domain-containing protein with MATH domain	3.91	0
AT5G52390	photoassimilate-responsive protein, putative.	3.91	0.86
AT4G17030	EXPANSIN-RELATED.	3.91	0
AT3G03660	Encodes a WUSCHEL-related homeobox gene family member.	4.27	0.16
AT4G34300	Encodes protein with 14.7% gly residues, similar to auxin response factor 30	4.27	0

**Table 3.1.** 33 genes upregulated both in light- and dark-grown DKO mutants

Overlap Genes	Description	FDR q-value (%)	
		Light	Dark
AT5G42600	Encodes oxidosqualene synthase that produces monocyclic triterpene marneral.	0	0
AT1G52050	jacalin lectin family protein.	0	0.16
AT2G16005	MD-2-related lipid recognition domain-containing protein with ML domain.	0	0
AT5G24140	Encodes a protein with similarity to squalene monooxygenases.	0	0.16
AT1G05650	polygalacturonase, putative / pectinase, putative.	0	0.16
AT1G77530	O-methyltransferase family 2 protein.	0	0
AT3G49190	condensation domain-containing protein	0	0.28
AT3G45680	proton-dependent oligopeptide transport (POT) family protein.	0	0
AT5G02000	unknown protein	0	0
AT5G35940	jacalin lectin family protein.	0	0
AT4G12545	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein.	0	0
AT5G04150	BHLH101	0	0.53
AT5G38020	encodes a protein like SAM:salicylic acid carboxyl methyltransferase (SAMT)	1.69	0
AT3G20940	a member of A-type cytochrome P450	1.69	0.28
AT5G47980	transferase family protein	1.69	0.28
AT1G73330	encodes a plant-specific protease inhibitor-like protein whose transcript level in root disappears in response to progressive drought stress.	1.69	0
AT2G25680	Encodes a high-affinity molybdate transporter.	1.69	0
AT3G06460	GNS1/SUR4 membrane family protein.	1.69	0
AT4G23590	aminotransferase class I and II family protein.	1.69	0
AT2G01520	MLP-LIKE PROTEIN 328 (MLP328)	1.69	0.28
AT1G77520	O-methyltransferase family 2 protein	1.69	1.19
AT2G37440	endonuclease/exonuclease/phosphatase family protein.	1.69	0
AT5G47450	Tonoplast intrinsic protein, transports ammonium (NH <sub>3</sub> ) and methylammonium.	1.69	0.28
AT4G22214	Encodes a defensin-like (DEFL) family protein.	1.69	0.28
AT1G79760	Identified as target of the AGL15 binding motif CArG.	1.69	0.28
AT4G22666	protease inhibitor/seed storage/lipid transfer protein (LTP)-like family protein.	1.69	0
AT5G46890	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein.	1.69	0
AT2G39040	peroxidase, putative; Identical to Peroxidase 24 precursor (PER24) [Arabidopsis Thaliana] (GB:Q9ZV04;GB:Q0V7W8).	1.69	0
AT1G63450	catalytic.	1.69	0
AT1G14960	major latex protein-related / MLP-related.	1.69	0.28
AT5G53190	nodulin MtN3 family protein.	1.85	0.28
AT4G26320	arabinogalactan protein 13 (AGP13)	2.37	2.55
AT4G25250	invertase/pectin methylesterase inhibitor family protein.	3.42	0.28
AT5G59090	ATSBT4.12; subtilase.	3.42	0
AT4G11210	disease resistance-responsive family protein / dirigent family protein.	3.42	0

AT1G51470	glycosyl hydrolase family 1 protein.	3.42	0
AT1G19900	glyoxal oxidase-related.	3.77	0
AT4G15290	encodes a gene similar to cellulose synthase	3.77	0
AT2G23410	encodes cis-prenyltransferase	3.77	0
AT4G13620	encodes member of DREB subfamily A-6 of ERF/AP2 transcription factor family.	3.77	0
AT4G14060	major latex protein-related / MLP-related	3.77	0.53
AT5G15600	SPIRAL1-LIKE4 belongs to a six-member gene family in Arabidopsis.	4.15	0
AT4G37160	SKS15 (SKU5 Similar 15); copper ion binding.	4.15	0
AT2G33790	pollen Ole e 1 allergen protein containing 14.6% proline residues.	4.15	0
AT1G05680	UDP-glucuronosyl/UDP-glucosyl transferase family protein.	4.15	0

**Table 3.2.** 45 genes down-regulated both in light- and dark-grown DKO mutants.

From Tables 3.1 and 3.2, we noticed that five up-regulated genes are stress related and three down-regulated genes can promote growth. A full list of these eight genes is given in Table 3.3.

		Up- or Down- regulated	FDR q-value (%)	
Gene	Description		Light	Dark
Stress-related				
AT4G33720	pathogenesis-related protein, putative.	Up	0.90	0
AT4G36430	peroxidase, putative.	Up	0.90	0
AT5G06720	peroxidase, putative.	Up	0.90	0
AT4G39830	L-ascorbate oxidase, putative.	Up	0.90	0
AT2G18150	peroxidase, putative.	Up	3.29	0
Growth-promoting				
AT4G15290	encodes a gene similar to cellulose synthase	Down	3.77	0
AT1G05680	UDP-glucuronosyl/UDP-glucosyl transferase family protein	Down	4.15	0
AT1G05650	polygalacturonase, putative / pectinase, putative	Down	0	0.16

**Table 3.3.** Five up-regulated genes are stress related and three down-regulated genes are growth-promoting genes.



As an independent confirmation of microarray data, real-time RT-PCR on five genes showing significant gene expression changes after the DKO treatment in dark-grown seedlings was performed. RNA was isolated in the same manner as samples used for microarray experiments. Table 3.4 shows that specific fold changes from real-time RT-PCR were reasonably close to those estimated by the microarray results.

Gene Locus	Up or Down-regulated	Fold Change	
		Microarray	Real-time RT-PCR
AT5G61890	Up	2.99	2.32
AT5G64060	Up	2.38	2.71
AT3G24650	No change	0.93	0.78
AT3G63110	Down	0.50	0.59
AT4G31320	Down	0.31	0.45

**Table 3.4.** Real-Time RT-PCR verification of microarray expression.

## DISCUSSION

In dark-grown seedlings, the most rapidly growing tissue is the hypocotyls whereas in light-grown seedlings the root is the most rapidly growing tissue. The growth of three transgenic lines was significantly reduced by the induction of an RNAi construct by estradiol in *apy2* mutants, both at the seedling and flowering stages of growth (Wu, Steinebrunner et al. 2007). In dark-grown seedlings, shortened hypocotyl length was found in 3.5-day-old etiolated seedlings of all three RNAi lines, while in light-grown seedlings, significantly shorter roots were observed in day 6 (Wu, Steinebrunner et al. 2007).

Comparing statistically significant genes that changed expression in the *apylapy2* double-knockout (DKO) mutants in both light- and dark-grown seedlings (Tables 3.1 and 3.2), we noticed that AT4G34300 was upregulated in both types of seedlings with FDR q-values of 4.27% in light and 0% in darkness. AT4G34300 encodes a protein with 14.7% glycine residues, similar to auxin response factor 30 (ARF30). By binding to auxin response elements on promoters of auxin response genes, Auxin response factors (ARF) are transcription factors that can either upregulate or downregulate the expression of auxin response genes (Guilfoyle and Hagen 2007). Tian et al. (2004) reported that the constitutive expression of ARF8 resulted in the suppression of the growth of Arabidopsis hypocotyls in the light, whereas the knockout of this gene resulted in the promotion of hypocotyl growth (Tian, Muto et al. 2004). Very little is known about ARF30, but given that its expression was enhanced when growth was suppressed by *APY* suppression, one might predict that ARF30 is like ARF8 in that its upregulation results in growth suppression.

One mechanism by which certain ARFs could suppress growth would be by repressing the expression of genes needed for auxin transport. In unpublished studies, Dr.

J. Wu in the Roux laboratory found that suppression of APY1 and APY2 resulted in plants with a dwarf phenotype and disrupted auxin distribution. One could speculate, then, AT4G34300 represses the expression of genes needed for auxin transport, and that its enhanced expression when *APY1* and *APY2* are suppressed leads to reduced auxin transport and reduced growth. This is an example of a testable hypothesis that arises from an analysis of the microarray data. It will be interesting to further investigate the involvement of AT4G34300 in plant growth and how apyrase suppression results in its upregulation.

Extracellular ATP (eATP) can act as a signaling molecule in the animal extracellular matrix by activating P2 receptor and subsequent downstream signal transduction cascades (Barnard, Simon et al. 1997). As a signaling agent, eATP may also do so in plant cells (Demidchik, Nichols et al. 2003). Previous studies have shown that cell release ATP as a consequence of growth (Kim, Sivaguru et al. 2006) and high levels of eATP can inhibit growth (Roux, Song et al. 2006) and the expression of APY1 and APY2 can lower the [eATP] of plant cells (Wu, Steinebrunner et al. 2007; Kim, Yang et al. 2009). In addition, Wu *et al.* (2007) proposed that plant cells may control their [eATP] to sustain growth and APY1 and APY2 may be key players in this mechanism.

A recent study (Kim, Yang et al. 2009) has shown that there is an increase of [eATP] due to hypertonic stresses, and that hypertonic stresses can induce the expression of APY1 and APY2. From Table 3.3, we noticed that five stress-related genes were upregulated and three growth-promoting genes are downregulated by the suppression of APY1 and APY2 in both light- and dark-grown seedlings. In addition, we found that five genes (AT2G02850, AT2G29460, AT5G59820, AT1G65500, AT5G20230) which were upregulated by the DKO treatment in dark-grown seedlings were also upregulated during RNA virus infection and significantly altered by CaLCuV infection (Ascencio-Ibanez,

Sozzani et al. 2008). These results corroborate the foregoing findings and hypotheses. Therefore, we propose that the suppression of APY1 and APY2 can induce the expression of stress-related genes. What do stress-related genes have to do with growth control?

When plants experience abiotic or biotic stress they undergo diverse physiological changes, one of which is typically growth reduction. That is, plants begin to reallocate the energy and other resources they use for protein production to begin making more of the proteins that protect them from the stress and less of the proteins needed for enhanced growth. Although blocking APY1 and APY2 production is not directly a biotic or abiotic stress, we could expect that some of the growth-regulating genes that change expression during plant stress responses would also be among the genes that help mediate the growth effects of apyrase suppression.

In addition to stress-related genes, Table 3.3 also highlighted some genes that the plant growth literature has implicated as needed to promote growth, consistent with their being down-regulated during growth inhibition. As yet we cannot identify which of these genes play key roles in mediating the growth changes that happen when plants become deficient in their production of APY1 and APY2. However, our NimbleGen results help us to focus on some likely candidates, and that is the key contribution that is provided by Tables 3.1, 3.2, and 3.3. Physiological and molecular experiments will be needed to confirm which of these candidates play important roles in apyrase-mediated growth changes. Related experiments to quantify how much [eATP] changes during growth and how increasing or decreasing the expression of APY1 and APY2 modulates these changes will also be needed.

## Chapter 4 Conclusions and Future Direction

The foregoing research work is mainly focused on microarray analysis of high-throughput biological data. In Chapter 2, we have developed a robust correlation coefficient, shrinkage correlation coefficient (SCC), which is an alternative to the Pearson correlation coefficient and the SD-weighted correlation coefficient, and particularly useful for clustering replicated microarray data generated by most academic laboratories. We have shown the superiority of SCC by the adjusted Rand index comparison on both synthetic and real expression data using hierarchical and k-means clustering. We apply SCC to successfully identify distinct clusters of genes during *C. richardii* early development. We also present the use of SVD to uncover the gene-wise biases introduced by experimental artifacts due to comparison of different biological replicates and prints of arrays. This computational approach is not only applicable to DNA microarray analysis but is also applicable to proteomics data or any other high-throughput analysis methodology.

The suppression of *APY1* and *APY2* in mutants expressing an inducible RNAi system resulted in plants with a dwarf phenotype and disrupted auxin distribution, and we used these mutants to discover what genes changed expression during growth suppression. A thorough analysis of the underlying gene expression changes associated with apyrase gene suppression was carried out in Chapter 3. We evaluated the gene expression changes of apyrase-suppressed RNAi mutants that had been grown in the light and in the darkness, using the NimbleGen *Arabidopsis thaliana* 4-Plex microarray, respectively. We compared the two sets of large-scale expression data and identified genes whose expression significantly changed after apyrase suppression in light and darkness, respectively. Our results allowed us to highlight some of the genes likely to

play major roles in mediating the growth changes that happen when plants drastically reduce their production of APY1 and APY2, some more associated with growth promotion and others, such as stress-induced genes, more associated with growth inhibition. There is a strong rationale for ranking all these genes as prime candidates for mediating the inhibitory growth effects of suppressing apyrase expression, thus the NimbleGen data will serve as a catalyst and valuable guide to the subsequent physiological and molecular experiments that will be needed to clarify the network of gene expression changes that accompany growth inhibition.

## **FUTURE DIRECTIONS**

In our previous study, we found that, in dark-grown seedlings, shortened hypocotyl length was found in 3.5-d-old etiolated seedlings of DKO mutants (Wu, Steinebrunner et al. 2007). However, from our recent data (unpublished), we noticed that, after the induction of the RNAi construct by estradiol in *apy2* mutants, the expression of *APY1* was significantly depressed since day 3. This suggests that some of the gene expression changes obtained from our dark-grown seedlings by NimbleGen microarray (3.5 day) are possibly secondary expression changes. Therefore, it is worth investigating the primary expression changes caused by the *apy1apy2* double-knockout (DKO) mutants (3 day) in dark-grown seedlings. As introduced in Chapter 1, RNA-Seq is “a revolutionary tool for transcriptomics” (Wang, Gerstein et al. 2009). We are planning to perform RNA-Seq on the 3-d DKO mutants to identify the primary gene expression changes caused by the suppression of APY1 and APY2. The combined microarray and RNA-Seq results will be very helpful to study the kinetics of gene expression after suppression of apyrases in *Arabidopsis*.

## Bibliography

- Alter, O., P. O. Brown, et al. (2000). "Singular value decomposition for genome-wide expression data processing and modeling." Proceedings of the National Academy of Sciences of the United States of America **97**(18): 10101-10106.
- Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis. New York, Wiley-Interscience.
- Ascencio-Ibanez, J. T., R. Sozzani, et al. (2008). "Global analysis of Arabidopsis gene expression uncovers a complex array of changes impacting pathogen response and cell cycle during geminivirus infection." Plant Physiology **148**(1): 436-454.
- Ashburner, M., C. A. Ball, et al. (2000). "Gene Ontology: tool for the unification of biology." Nature Genetics **25**(1): 25-29.
- Banks, J. A., L. Hickok, et al. (1993). "The Programming Of Sexual Phenotype In The Homosporous Fern *Ceratopteris-Richardii*." International Journal Of Plant Sciences **154**(4): 522-534.
- Barnard, E. A., J. Simon, et al. (1997). "Nucleotide receptors in the nervous system - An abundant component using diverse transduction mechanisms." Molecular Neurobiology **15**(2): 103-129.
- Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society Series B-Methodological **57**(1): 289-300.
- Benjamini, Y. and D. Yekutieli (2001). "The control of the false discovery rate in multiple testing under dependency." Annals of Statistics **29**(4): 1165-1188.
- Bentley, D. R. (2006). "Whole-genome resequencing." Current Opinion in Genetics & Development **16**(6): 545-552.
- Bland, M. (1995). An Introduction to Medical Statistics Oxford University Press.
- Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di test. Studi in onore del Professore Salvatore Ortu Carboni. Rome: 13-60.
- Bove, J., P. Lucas, et al. (2005). "Gene expression analysis by cDNA-AFLP highlights a set of new signaling networks and translational control during seed dormancy breaking in *Nicotiana plumbaginifolia*." Plant Molecular Biology **57**(4): 593-612.



Chatterjee, A., D. M. Porterfield, et al. (2000). "Gravity-directed calcium current in germinating spores of *Ceratopteris richardii*." Planta **210**(4): 607-610.

Chatterjee, A. and S. J. Roux (2000). "Ceratopteris richardii: A productive model for revealing secrets of signaling and development." Journal Of Plant Growth Regulation **19**(3): 284-289.

Cheremsky, V., J. Feng, et al. (2003). "Shrinkage-based similarity metric for cluster analysis of microarray data." Proceedings of the National Academy of Sciences of the United States of America **100**(17): 9668-9673.

Cloonan, N., A. R. R. Forrest, et al. (2008). "Stem cell transcriptome profiling via massive-scale mRNA sequencing." Nature Methods **5**(7): 613-619.

Cui, X. G., J. T. G. Hwang, et al. (2005). "Improved statistical tests for differential gene expression by shrinking variance components estimates." Biostatistics **6**(1): 59-75.

D'Haeseleer, P. (2005). "How does gene expression clustering work?" Nature Biotechnology **23**(12): 1499-1501.

Decker, E. L., W. Frank, et al. (2006). "Moss systems biology en route: Phytohormones in *Physcomitrella* development." Plant Biology **8**(3): 397-405.

Demeter, J., C. Beauheim, et al. (2007). "The Stanford Microarray Database: implementation of new analysis tools and open source release of software." Nucleic Acids Research **35**: D766-D770.

Demidchik, V., C. Nichols, et al. (2003). "Is ATP a signaling agent in plants?" Plant Physiology **133**(2): 456-461.

Draghici, S. (2003). Data Analysis Tools for DNA Microarrays, Chapman & Hall/CRC.

Efron, B. and C. Morris (1973). "Stein's Estimation Rule and Its Competitors--An Empirical Bayes Approach " Journal of the American Statistical Association: 117-130.

Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proceedings of the National Academy of Sciences of the United States of America **95**(25): 14863-14868.

Furuhashi, K. (2002). "Involvement of actin dephosphorylation in germination of *Physarum sclerotium*." Journal Of Eukaryotic Microbiology **49**(2): 129-133.

Golub, G. H. and C. F. Van Loan (1996). Matrix Computations. Baltimore, The Johns Hopkins University Press.

Guilfoyle, T. J. and G. Hagen (2007). "Auxin response factors." Current Opinion in Plant Biology **10**(5): 453-460.

Gulemetova, R., H. Chamberland, et al. (1998). "Presence of small-nuclear-ribonucleoprotein-containing nuclear bodies in quiescent and early germinating Zea mays embryos." Protoplasma **202**(3-4): 192-201.

Hartigan, J. A. (1975). Clustering Algorithms. New York, John Wiley and Sons.

Holter, N. S., M. Mitra, et al. (2000). "Fundamental patterns underlying gene expression profiles: Simplicity from complexity." Proceedings of the National Academy of Sciences of the United States of America **97**(15): 8409-8414.

Hubert, L. and P. Arabie (1985). "Comparing Partitions." Journal Of Classification **2**(2-3): 193-218.

Hughes, T. R., M. J. Marton, et al. (2000). "Functional discovery via a compendium of expression profiles." Cell **102**(1): 109-126.

Ideker, T., V. Thorsson, et al. (2001). "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network." Science **292**: 929-934.

James, W. and C. Stein (1961). Estimation with quadratic loss. Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, Berkeley, University of California Press.

Kasturi, J., R. Acharya, et al. (2003). "An information theoretic approach for analyzing temporal patterns of gene expression." Bioinformatics **19**(4): 449-458.

Kermode, A. R. (2005). "Role of abscisic acid in seed dormancy." Journal Of Plant Growth Regulation **24**(4): 319-344.

Kerr, M. K. and G. A. Churchill (2001). "Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments." Proceedings of the National Academy of Sciences of the United States of America **98**(16): 8961-8965.

Killion, P., G. Sherlock, et al. (2003). "The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD)." BMC Bioinformatics **4**: 32.

- Kim, S. H., S. H. Yang, et al. (2009). "Hypertonic Stress Increased Extracellular ATP Levels and the Expression of Stress-Responsive Genes in *Arabidopsis thaliana* Seedlings." Bioscience Biotechnology and Biochemistry **73**(6): 1252-1256.
- Kim, S. Y., M. Sivaguru, et al. (2006). "Extracellular ATP in plants. Visualization, localization, and analysis of physiological significance in growth and signaling." Plant Physiology **142**(3): 984-992.
- Kishi, Y., D. Mahadeo, et al. (2000). "Glucose-induced pathways for actin tyrosine dephosphorylation during *Dictyostelium* spore germination." Experimental Cell Research **261**(1): 187-198.
- Klevecz, R. R., J. Bolen, et al. (2004). "From the Cover: A genomewide oscillation in transcription gates DNA replication and cell cycle." Proceedings of the National Academy of Sciences of the United States of America **101**(5): 1200-1205.
- Kung, C., D. M. Kenski, et al. (2005). "Chemical genomic profiling to identify intracellular targets of a multiplex kinase inhibitor." Proceedings of the National Academy of Sciences of the United States of America **102**(10): 3587-3592.
- Langmead, B., C. Trapnell, et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biology **10**(3): 10.
- Ledoit, O. and M. Wolf (2004). "A well-conditioned estimator for large-dimensional covariance matrices." Journal of multivariate analysis **88**: 365-411.
- Lee, H. K., W. Braynen, et al. (2005). "ErmineJ: Tool for functional analysis of gene expression data sets." BMC Bioinformatics **6**: 269.
- Li, C. M. and R. R. Klevecz (2006). "From the Cover: A rapid genome-scale response of the transcriptional oscillator to perturbation reveals a period-doubling path to phenotypic change." Proceedings of the National Academy of Sciences of the United States of America **103**(44): 16254-16259.
- Li, H., J. Ruan, et al. (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome REsearch **18**: 1851-1858.
- Li, R., Y. Li, et al. (2008). "SOAP: short oligonucleotide alignment program." Bioinformatics **24**: 713-714.
- Lin, H., Z. Zhang, et al. (2008). "ZOOM! Zillions Of Oligos Mapped." Bioinformatics **24**: 2431-2437.

Lister, R., R. C. O'Malley, et al. (2008). "Highly integrated single-base resolution maps of the epigenome in Arabidopsis." Cell **133**(3): 523-536.

Lorenzo, O., C. Nicolas, et al. (2002). "Molecular cloning of a functional protein phosphatase 2C (FsPP2C2) with unusual features and synergistically up-regulated by ABA and calcium. in dormant seeds of Fagus sylvatica." Physiologia Plantarum **114**(3): 482-490.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, University of California Press.

Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Marioni, J. C., C. E. Mason, et al. (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." Genome Research **18**(9): 1509-1517.

Matsumura, H., K. H. Bin Nasir, et al. (2006). "SuperSAGE array: the direct use of 26-base-pair transcript tags in oligonucleotide arrays." Nature Methods **3**(6): 469-474.

McShane, L. M., M. D. Radmacher, et al. (2002). "Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data." Bioinformatics **18**(11): 1462-1469.

Medvedovic, M., K. Y. Yeung, et al. (2004). "Bayesian mixture model based clustering of replicated microarray data." Bioinformatics **20**(8): 1222-1232.

Milligan, G. W. and M. C. Cooper (1986). "A Study Of The Comparability Of External Criteria For Hierarchical Cluster-Analysis." Multivariate Behavioral Research **21**(4): 441-458.

Monti, S., K. J. Savage, et al. (2005). "Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response." Blood **105**(5): 1851-1861.

Morin, R., M. Bainbridge, et al. (2008). "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing." Biotechniques **45**(1): 81-94.

Mortazavi, A., B. A. Williams, et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nature Methods **5**(7): 621-628.

- Nagalakshmi, U., Z. Wang, et al. (2008). "The transcriptional landscape of the yeast genome defined by RNA sequencing." Science **320**(5881): 1344-1349.
- Ng, S. K., G. J. McLachlan, et al. (2006). "A Mixture model with random-effects components for clustering correlated gene-expression profiles." Bioinformatics **22**(14): 1745-1752.
- Nielsen, T. O., R. B. West, et al. (2002). "Molecular characterisation of soft tissue tumours: a gene expression study." Lancet **359**(9314): 1301-1307.
- Ogawa, M., A. Hanada, et al. (2003). "Gibberellin biosynthesis and response during Arabidopsis seed germination." Plant Cell **15**(7): 1591-1604.
- Rengarajan, J., B. R. Bloom, et al. (2005). "From The Cover: Genome-wide requirements for Mycobacterium tuberculosis adaptation and survival in macrophages." Proceedings of the National Academy of Sciences of the United States of America **102**(23): 8327-8332.
- Reyes, D., D. Rodriguez, et al. (2006). "Evidence of a role for tyrosine dephosphorylation in the control of postgermination arrest of development by abscisic acid in Arabidopsis thaliana L." Planta **223**(2): 381-385.
- Roux, S. J., C. Song, et al. (2006). Regulation of plant growth and development by extracellular nucleotides. Communication in Plants. F. Baluska, S. Mancuso and D. Volkmann. Berlin, Springer Berlin Heidelberg: 221-234.
- Rumble, S. M., P. Lacroute, et al. (2009). "SHRiMP: Accurate Mapping of Short Color-space Reads." Plos Computational Biology **5**(5): 11.
- Saez, A., N. Robert, et al. (2006). "Enhancement of abscisic acid sensitivity and reduction of water consumption in Arabidopsis by combined inactivation of the protein phosphatases type 2C ABI1 and HAB1." Plant Physiology **141**(4): 1389-1399.
- Salmi, M. L., T. J. Bushart, et al. (2005). "Profile and analysis of gene expression changes during early development in germinating spores of *Ceratopteris richardii*." Plant Physiology **138**(3): 1734-1745.
- Schäfer, J. and K. Strimmer (2005). "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics." Statistical Applications in Genetics and Molecular Biology **4**: Article 32.
- Schnepf, E. and C. Reinhard (1997). "Brachycytes in Funaria protonemate: Induction by abscisic acid and fine structure." Journal Of Plant Physiology **151**(2): 166-175.

Shaw, B. D. and S. Upadhyay (2005). "Aspergillus nidulans swoK encodes an RNA binding protein that is important for cell polarity." Fungal Genetics And Biology **42**(10): 862-872.

Shen, Y. Y., X. F. Wang, et al. (2006). "The Mg-chelatase H subunit is an abscisic acid receptor." Nature **443**(7113): 823-826.

Sidak, Z. (1967). "Rectangular confidence regions for the means of multivariate normal distributions." Journal of the American Statistical Association **62**: 626-633.

Smith, A. D., Z. Xuan, et al. (2008). "Using quality scores and longer reads improves accuracy of Solexa read mapping." BMC Bioinformatics **9**: 128.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. Proc. Third. Berkeley Symp. Math. Statist. Probab., Berkeley, Univ. California Press.

Steinebrunner, I., J. Wu, et al. (2003). "Disruption of apyrases inhibits pollen germination in arabidopsis." Plant Physiology **131**(4): 1638-1647.

Storey, J. D. and R. Tibshirani (2003). "Statistical significance for genomewide studies." Proceedings of the National Academy of Sciences of the United States of America **100**(16): 9440-9445.

Tian, C., H. Muto, et al. (2004). "Disruption and overexpression of auxin response factor 8 gene of Arabidopsis affect hypocotyl elongation and root growth habit, indicating its possible involvement in auxin homeostasis in light condition." Plant Journal **40**(3): 333-343.

Tjaden, B. (2006). "An approach for clustering gene expression data with error information." Bmc Bioinformatics **7**.

Troyanskaya, O., M. Cantor, et al. (2001). "Missing value estimation methods for DNA microarrays." Bioinformatics **17**: 520-525.

Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proceedings of the National Academy of Sciences of the United States of America **98**(9): 5116-5121.

van't Veer, L. J., H. Y. Dai, et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer." Nature **415**(6871): 530-536.

Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature Reviews Genetics **10**(1): 57-63.

Wilhelm, B. T. and J. R. Landry (2009). "RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing." Methods **48**(3): 249-257.

Wilhelm, B. T., S. Marguerat, et al. (2008). "Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution." Nature **453**(7199): 1239-U39.

Wu, J., I. Steinebrunner, et al. (2007). "Apyrases (nucleoside triphosphate-diphosphohydrolases) play a key role in growth control in arabidopsis." Plant Physiology **144**(2): 961-975.

Yeung, K. Y., D. R. Haynor, et al. (2001). "Validating clustering for gene expression data." Bioinformatics **17**(4): 309-318.

Yeung, K. Y., M. Medvedovic, et al. (2003). "Clustering gene-expression data with repeated measurements." Genome Biology **4**(5): R 34.

Yeung, M. K. S., J. Tegner, et al. (2002). "Reverse engineering gene networks using singular value decomposition and robust regression." Proceedings of the National Academy of Sciences of the United States of America **99**(9): 6163-6168.

## **Vita**

Jianchao Yao was born in Tongling, Anhui, China on March 20, 1978, the son of Aiping He and Xueyi Yao. In 1996, he graduated from Tongling First High School in Tongling, Anhui, China. He received his Bachelor of Engineering degree in Biomedical Engineering from Southeast University, Nanjing, China in June, 2000. He then entered the University of Hong Kong and received a Master of Philosophy degree in Biomedical Engineering in August, 2003. In the fall semester of 2003, he was enrolled in the Graduate School of the University of Texas at Austin as a graduate student in the Track of Bioinformatics & Computational Biology in the Cell and Molecular Biology Program. While working toward his Ph.D., he also studied in the Master's in Statistics Program at the University of Texas at Austin and received a Master of Science degree in Statistics in August, 2007.

Permanent address: Room 509, Building 9, Changjiang Dong Cun  
Tongling, Anhui 244000  
China

This dissertation was typed by the author.